

Constructing A Validity Argument for A Locally Developed Test of English Reading Proficiency

Vo Ngoc Hoi

BA (English Teacher Education)

MA (Teaching English to Speakers of Other Languages)

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

School of Teacher Education and Leadership
Faculty of Education
Queensland University of Technology

2021

Keywords

L-VSTEP reading test, construct validity, argument-based validation, explanation inference, extrapolation inference, reading processes, predictors of item difficulty, factor structure, factorial invariance, self-assessment, construct validity, predictive validity, mixed methods research, Vietnam

Abstract

This project aims to build a validity argument for the reading component of a locally developed Vietnamese Standardized Test of English Proficiency (L-VSTEP) – a new high-stakes test used by Vietnamese universities as evidence of graduates’ English proficiency. A mixed-method paradigm is employed to examine three interrelated aspects of the construct of the L-VSTEP reading test, thereby offering insights into the extent to which the pattern of test scores, the reading processes of test takers, and the linguistic features of the reading texts correspond with what is intended to be measured by the test. The relationship between students’ scores on the test and their performance in the academic domains at the relevant tertiary institution is also investigated to shed light on the meaning of the test scores beyond the test per se.

Results of the study provide some evidence in support of the interpretation and use of the test scores. Supporting evidence includes the similarity between expert judgment and students’ verbal reports in terms of the reading skills elicited by the test items, the identifiable factor structure of the test as aligned with the guidelines for test item writing, the predictive relationship between students’ scores on the test and their self-reported reading performance in the target language use domain, and the alignment between the reading skills assessed in the test and those encountered in the target language use domain, particularly for English major students.

The study also generates some rebuttals that might weaken the validity argument of the test score interpretation and use. Rebutting evidence entails the limited range of reading skills among low-achieving students that the test can assess, students’ use of test taking strategies that might introduce construct-irrelevant variance, linguistic features of the test questions overshadowing those of the texts in predicting item difficulty, non-invariance of the factor structure of the test across students with different academic disciplines and reading performance levels, and misalignment of the reading tasks and skills assessed in the test and those encountered in the target language use domain, particularly for non-English major students.

The study offers potential implications for relevant stakeholders of the test including test designers, policy makers, teachers, students, researchers, and curriculum designers, and contributes to a growing body of research that adopts an argument-based approach to language test validation.

Table of Contents

Keywords	i
Abstract	ii
Table of Contents	iii
List of Tables.....	vii
List of Figures	ix
List of Abbreviations.....	x
Statement of Original Authorship	xi
Acknowledgement.....	xii
CHAPTER I: INTRODUCTION	1
1.1. Background of the study	1
1.1.1. The global context	1
1.1.2. The local context	2
1.2. Research context and research problems	3
1.3. Definitions of key concepts.....	6
1.4. Significance of the study	10
1.5. Thesis structure	10
CHAPTER II: LITERATURE REVIEW	12
2.1. The nature of L2 reading comprehension	13
2.2. Constructs of L2 reading.....	16
2.2.1. The processing perspective.....	16
2.2.2. The task perspective	23
2.2.3. The reader purpose perspectives	30
2.3. Chapter summary	35
CHAPTER III: THEORETICAL FRAMEWORK	37
3.1. Toulmin’s argument structure	37
3.2. The argument-based approach to test validation.....	39
3.2.1. The interpretive argument	39
3.2.2. The validity argument.....	43
3.2.3. Implementation of the argument-based approach	45
3.3. Articulating an interpretive argument for the L-VSTEP reading test	48
3.3.1. Description of the L-VSTEP reading test.....	48
3.3.2. The explanation inference	50
3.3.3. The extrapolation inference	56

CHAPTER IV: METHODOLOGY	59
4.1. Mixed methods research	59
4.2. Participants.....	63
4.2.1. Sampling methods and participant recruitment.....	63
4.2.2. Sample size requirements	67
4.3. Data collection and data analysis procedure	69
4.3.1. Research question 1	69
4.3.2. Research question 2.....	81
4.3.3. Research question 3	93
4.3.4. Research question 4.....	107
4.3.5. Research question 5	111
CHAPTER V: EXPERT JUDGMENT AND STUDENTS' REPORTED READING PROCESSES	117
5.1. Introduction.....	117
5.2. Findings from expert judgment.....	117
5.2.1. Findings from the pilot stage.....	117
5.2.2. Findings from the main stage	120
5.2.3. Discussion of the expert judgment findings	124
5.3. Findings from students' verbal reports	125
5.3.1. Items mainly assessing Understanding Explicit Information.....	126
5.3.2. Items mainly assessing Lexical Inferencing.....	129
5.3.3. Items mainly assessing Understanding Cohesive Devices.....	133
5.3.4. Items mainly assessing Integrating Textual Information	137
5.3.5. Items assessing Inferring Situational Meaning.....	141
5.3.6. Items assessing Understanding Pragmatic Meaning	145
5.3.7. Items mainly assessing Summarizing Textual Information	148
5.3.8. Items assessing Identifying Text Structure.....	152
5.4. Discussion	152
CHAPTER VI: FACTOR STRUCTURE AND FACTORIAL INVARIANCE OF THE TEST	158
6.1. Introduction.....	158
6.2. Item coding and unit of measurement for Confirmatory Factor Analysis (CFA).....	158
6.3. CFA model building.....	161
6.4. Results.....	164

6.4.1. Descriptive statistics	164
6.4.2. CFA findings	171
6.4.3. Factorial invariance of the one-factor model.....	178
6.5. Discussion	182
CHAPTER VII: TEXT AND ITEM FEATURES AS PREDICTORS OF ITEM DIFFICULTY.....	187
7.1. Introduction	187
7.2. Descriptive statistics.....	187
7.2.1. Linguistic and discourse features of reading texts	187
7.2.2. Item difficulty via Rasch modeling	189
7.3. Findings of the correlational and multiple regression analyses	190
7.3.1. Correlation analysis	190
7.3.2. Regression analyses.....	192
7.4. Discussion	198
CHAPTER VIII: STUDENTS' TEST SCORES AND THEIR ENGLISH READING PROFICIENCY IN THE TARGET LANGUAGE USE DOMAIN.....	203
8.1. Introduction	203
8.2. Validation of the self-assessment instrument.....	203
8.2.1. Exploratory factor analysis.....	203
8.2.2. Confirmatory factor analysis	208
8.3. Test performance and self-reported English reading proficiency	215
8.4. Discussion	219
CHAPTER IX: ALIGNMENT OF THE L-VSTEP READING TEST TO THE TARGET LANGUAGE USE DOMAINS	224
9.1. Introduction	224
9.2. Findings.....	224
9.2.1. The perceived importance of English reading in the academic programs.....	225
9.2.2. The amount and type of English reading required in the academic programs ...	227
9.2.3. The reading tasks and skills required in the academic domains.....	230
9.2.4. Comparability between the academic domains and the L-VSTEP reading test.	235
9.3. Discussion	240
CHAPTER X: VALIDITY ARGUMENT FOR THE L-VSTEP READING TEST	243
10.1. Introduction	243
10.2. The explanation inference	247

10.3. The extrapolation inference.....	251
CHAPTER XI: CONCLUSION	255
11.1. Summary of the research findings.....	255
11.2. Theoretical implications.....	255
11.3. Methodological implications.....	256
11.4. Practical implications.....	257
11.5. Limitations	261
11.6. Future directions.....	262
REFERENCES.....	264
APPENDICES	280

List of Tables

Table 3. 1. Score range for the VSTEP reading test	48
Table 3. 2. Description of reading proficiency based on L-VSTEP reading test scores.....	49
Table 3. 3. Summary of the explanation inference for the L-VSTEP reading test.....	54
Table 3. 4. Summary of the extrapolation inference for the L-VSTEP reading test.....	57
Table 4. 1. The convergent parallel mixed methods design	62
Table 4. 2. Sampling of participants in the project	66
Table 4. 3. Textual features of the L-VSTEP reading test Form A	72
Table 4. 4. Background information of the participants	78
Table 4. 5. Distribution of the study participants.....	96
Table 4. 6. Summary of text, item and item-text variables.....	98
Table 4. 7. Interview questions	114
Table 5. 1. Reading skills, definitions and descriptions	118
Table 5. 2. Results of the expert judgment of reading skills.....	120
Table 5. 3. Findings from students' protocols on the Understanding Explicit Information subskill	127
Table 5. 4. Findings from students' verbal protocols on the Lexical inferencing subskill....	130
Table 5. 5. Findings from students' verbal protocols on the Understanding Cohesive Devices subskill	134
Table 5. 6. Findings from students' verbal protocols on the Integrating Textual Information subskill	139
Table 5. 7. Findings from students' verbal protocols on the Inferring Situational Meaning subskill	143
Table 5. 8. Findings from students' verbal protocols on the Understanding Pragmatic Meaning subskill.....	147
Table 5. 9. Findings from students' verbal protocols on the Summarizing Textual Information subskill	149
Table 6. 1. Expert judgment of reading subskills and the relevant test items.....	159
Table 6. 2. Descriptive statistics (N = 544)	165
Table 6. 3. Separation, reliability and unidimensionality measures	168
Table 6. 4. Dimensionality of the item parcels	169
Table 6. 5. Descriptive statistics of the parcels (N = 544).....	170
Table 6. 6. The global model fit indices	171

Table 6. 7. Unstandardized parameter estimates	172
Table 6. 8. The unstandardized estimates of the three correlated factor model.....	174
Table 6. 9. The unstandardized parameter estimates of the second-order factor model.....	175
Table 6. 10. Descriptive statistics of the sub-groups	179
Table 6. 11. The metric models	180
Table 6. 12. Goodness-of-fit indices for the low-scoring and high-scoring groups	182
Table 7. 1. Text and item features.....	188
Table 7. 2. Descriptive statistics of the item properties.....	189
Table 7. 3: Correlation between text, item, and item-text variables and item difficulty	190
Table 7. 4. Model summary	194
Table 7. 5. Parameter estimates	196
Table 7. 6. Hierarchical model summary.....	196
Table 7. 7. Parameter estimates of the hierarchical regression model.....	197
Table 8. 1. Descriptive statistics of the self-assessment questionnaire (N = 344).....	204
Table 8. 2. Factor solution and factor loadings.....	206
Table 8. 3. Goodness-of-fit indices for the three CFA models.....	209
Table 8. 4. Unstandardized parameter estimates of the second-order factor model.....	213
Table 8. 5. Parameter estimates of the test model.....	215
Table 8. 6. The unstandardized parameter estimates of the SEM model.....	217
Table 10. 1. The validity argument for the L-VSTEP reading test.....	244

List of Figures

Figure 2. 1. The integrated model of L2 reading	12
Figure 2. 2. The cognitive processing model (Brunfaut & McCray, 2015b).....	21
Figure 3. 1. Toulmin’s Argument Structure.....	38
Figure 3. 2. The interpretive argument	40
Figure 3. 3. The three-plane model of explanation inference (Chapelle et al., 2008, p.336) ..	53
Figure 4. 1. Data analysis procedure for RQ1	71
Figure 4. 2. Procedure for collecting and analysing stimulated verbal recall data	77
Figure 4. 3. Data analysis procedure for RQ ₂	85
Figure 4. 4. Data analysis procedure for RQ ₃	95
Figure 4. 5. The data analysis procedure for RQ ₄	108
Figure 4. 6. The data analysis procedure for RQ ₅	112
Figure 6. 1. The general reading proficiency model	162
Figure 6. 2. The correlated three factor model.....	163
Figure 6. 3. The higher-order factor model.....	164
Figure 6. 4. The Wright map (N = 544).....	167
Figure 6. 5. The one-factor model with standardized estimates	173
Figure 6. 6. The three correlated factor model with standardized estimates	175
Figure 6. 7. The second-order factor model with standardized estimates	177
Figure 6. 8. The metric model with equality constraints	181
Figure 7. 1. The scatter plot of standardized residuals	193
Figure 7. 2. The P-P plot of standardized residuals	194
Figure 8. 1. The scree plot	206
Figure 8. 2. The hypothesized CFA models	209
Figure 8. 3. The standardized parameter estimates of the correlated factor model	211
Figure 8. 4. Standardized parameters of the second order factor model.....	212
Figure 8. 5. The structural equation model.....	216

List of Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CEFR	Common European Framework of References for Languages
CEFR-VN	Vietnamese Version of the Common European Framework of References
EFL	English as A Foreign Language
L1	First Language
L2	Second Language
L-VSTEP	The Locally Developed Vietnamese Standardized Test of English Proficiency
MNSQ	Mean Square
PTMEA	Point Measure Correlation
RMSEA	Root Mean Square Error of Approximation
SEM	Structural Equation Modeling
SRMR	Standardized Root Mean Square Residual
TLI	Tucker Lewis Index
UEI	Understanding Explicit Meaning
UPM	Understanding Pragmatic Meaning
UCD	Understanding Cohesive Devices
ISM	Inferring Situational Meaning
STI	Summarizing Textual Information
ITI	Integrating Textual Information

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no materials previously published or written by another person except where due reference is made

Signature: [QUT Verified Signature](#)

Date: 16/02/2021

Acknowledgement

I am not alone on the journey to the completion of this thesis. There are many people around who have guided, supported, challenged, and encouraged me to reap the full benefits that this PhD journey has to offer. I would like to take this opportunity to acknowledge these people.

First and foremost, I am deeply grateful to my supervisory team, Dr. Lyn May and Associate Professor Guanglun Michael Mu who have introduced me to the world of academia, guided me through the ups and downs of this intellectual journey, cheered me up when I was down, challenged me to expand my knowledge boundaries, and encouraged me to become better academically, professionally, and personally. This thesis would not have been possible without their unwavering support, mentorship, and encouragement. Secondly, I would like to express my gratitude to all the participants who have volunteered their time and efforts to complete the tests, the questionnaire, and the interviews that constitute the major sources of data for my study. They have contributed significantly to the completion of this thesis. Thirdly, I am grateful to the panel members who have provided constructive feedback and suggestions on my document at the confirmation and final oral stages, Professor Greg Thompson, Dr. Chris Blundell, and Dr. Radha Iyer. Fourthly, I would like to thank the staff in the Education Research Office, Faculty of Education, Queensland University of Technology for their timely assistance throughout my PhD journey. Special thanks also go to my PhD fellows who have accompanied me along the way, gave me advice, and cheered for my milestones and achievements. Finally, I am deeply indebted to my beloved family for their endless love, care, encouragement, and sacrifice. This thesis is dedicated to them.

This thesis was conducted under the financial sponsorship of Queensland University of Technology via the Queensland University of Technology Post Graduate Research Award. I would like to thank the university for giving me this life-changing opportunity.

CHAPTER I: INTRODUCTION

This study aims to construct a validity argument for the reading component of a locally developed Vietnamese Standardized Test of English Proficiency (L-VSTEP), a high-stakes test of English proficiency first introduced in 2015 for various purposes including university exit requirements in Vietnam. This study is essential and timely given that the L-VSTEP is a newly developed and institutionalized high-stakes test of English proficiency affecting various stakeholders including, but not limited to, students, teachers, policy makers, and employers. In addition, there is currently little published empirical evidence regarding the interpretation and use of the test scores. This chapter introduces the background of the study, followed by a discussion of the key concepts, research aims and significance as well as the research questions to be explored.

1.1. Background of the study

1.1.1. The global context

The field of language testing and assessment has a long history and a far reaching impact on different aspects of language use and language teaching and learning. People engage in language testing and assessment for a variety of different purposes, from the use of language testing in the wider world such as ensuring effective communication in air traffic control, or informing government immigration policy, to the use of language assessment in the classroom to facilitate teaching and learning English, or to serve as a gate-keeping tool for university admission and graduation (Bachman & Palmer, 2010; Fulcher & Davidson, 2012). Concerning the latter, Green (2014) distinguished educational assessment from proficiency assessment. Educational assessment is closely related to the learning of languages in classroom contexts where assessment is used to support and document learning progress, to align the learning and teaching practice to educational goals, and to inform the selection of instructional materials germane to learners of different ability levels. On the other hand, proficiency assessment shifts the focus from examining what the language learners have been taught in a language program to what they can do with the repertoire of language knowledge and skills they have at their disposal, thereby determining the adequacy of their language ability vis-a-vis a predetermined standard or criterion (Green, 2014, p.13).

Language proficiency assessments can be used to make a range of high-stakes decisions. These include, but are not limited to, whether a student can meet the linguistic demands in an academic program where the target language constitutes the medium of instruction, whether a job candidate can communicate efficiently in a workplace where the language of communication is not his/her first language, or whether a person can be granted permanent residency status in another country. Examples of proficiency language assessments include established international standardized English proficiency tests such as the International English Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL), and a growing number of national and regional English proficiency tests, such as the General English Proficiency Test (GEPT) (Taiwan), the College English Test (CET) (China), and most recently the VSTEP (Vietnam).

High-stakes English language tests are taken by millions of students each year and are used by numerous institutions, societies, and government agencies to make decisions for employment and recruitment purposes. The interpretation and use of the scores of these tests, and the decisions made upon them have profound consequences for various stakeholders including those who provide resources for the development of the tests as well as those who are affected by the tests, such as students, teachers, and policy makers (Bachman & Palmer, 2010). It is, therefore, the responsibility of anyone who is involved with a specific test, and with the language testing and assessment practice at large, to make sure that the intended interpretations and uses of their test are justifiable on the basis of theoretical and empirical evidence, and to hold users of their test accountable for understanding how to interpret, use, and justify the test scores by giving them essential information and instructional materials (International Language Testing Association, 2007). This is where language test validation comes into play and where the current study is conceptually situated.

1.1.2. The local context

Socio-economic development and wider global integration in recent years have driven the transformation of education in Vietnam. Particularly, tertiary education has undergone extensive modifications in adaptation to the ever changing human resource demand of a growing global market (British Council, 2017). In this context, English proficiency has been assumed to be a necessary tool for nation building and an essential passport to the world of international universities and professional opportunities for students (Le, 2017). In keeping with the increasingly important role of English as a gatekeeper for educational advancement and

occupational development, the learning and teaching of English in Vietnam have been prioritized over the past decades. Numerous attempts have been made to raise the standard of English proficiency of both teachers and students at different levels of the educational system through various projects, reforms, and new policies, the current focus of which is on the National Foreign Language Project 2020 (hereafter Project 2020) (Government of Vietnam (GoV), 2008).

The Project 2020 aims to adopt international benchmarking standards for enhancing English proficiency of both teachers and students at all levels (Le, 2017). This is to be achieved through four initiatives: making English a compulsory subject at all educational levels, improving teachers' English proficiency and pedagogical skills through various teacher training courses, designing new curriculum and English materials, and standardizing teachers' and students' English proficiency by introducing the Vietnamese version of the Common European Framework of References for languages (CEFR-VN) which is a close adaptation of the Common European Framework of References (CEFR) level and level descriptors (Council of Europe, 2001) for use in the context of Vietnam (GoV, 2008; Le, 2017; Ministry of Education and Training of Vietnam (MOET), 2014; Tannenbaum & Baron, 2015). The CEFR-VN is a 6-level national language proficiency scale used as a benchmark for language training programs, language training institutions, and learners' language proficiency across the educational systems. The CEFR-VN is composed of six levels (1, 2, 3, 4, 5, 6) corresponding with the six levels of the CEFR (A1, A2, B1, B2, C1, C2), each of which is described in the document in sufficient detail so that its application at local levels can be easily carried out (MOET, 2014).

Another important task of the Project 2020 is the development of a national standardized English proficiency test that is aligned with the CEFR-VN and that can be used to assess English learners' proficiency levels at a nationwide level. This culminated in the introduction of the Vietnamese three-level test of English proficiency targeted at Levels 3 to 5 of the CEFR-VN in early 2015 (Dunlea et al., 2018; MOET, 2015b). Another test development project which is not the focus of this study and which embraces Levels 1 and 2 of the CEFR-VN is currently underway. The next section offers a brief overview of the Vietnamese three-level test of English proficiency, the research context and research problems to be addressed in this thesis.

1.2. Research context and research problems

The introduction of the Vietnamese three-level test of English proficiency (hereafter, the VSTEP) as a standardised test of English proficiency for Vietnamese adult learners of English

resulted from the concerted effort of a group of experts from the University of Languages and International Studies (ULIS) mandated by the Ministry of Education and Training (MOET) of Vietnam and several international language testing experts (Dunlea et al., 2018; Nguyen, 2020). The MOET introduced the VSTEP test in 2015 through a series of guidelines stated in Decree 729/QĐ-BGDĐT (MOET, 2015b). The test is mapped onto the CEFR-VN at Level 3/B1 to Level 5/C1 and is intended to serve a variety of purposes including university graduation, professional accreditation, and academic promotion. Since its introduction, many training workshops for item writers, teachers, and raters have been offered nationwide. Fifteen test sites are now officially authorized to organize the test on a regular basis at the national level, while other tertiary institutions are allowed to design and use in-house tests for their own internal purposes. Numerous training workshops and seminars have been organized nationwide to familiarize test-takers, teachers, test designers, and raters with the format, rating scale, requirements, and test item writing procedures of the test. As described by Le (2017, p.187),

“ ... the VSTEP is developed on the basis of test format and test specifications which have been verified in terms of validity and reliability. The scores of the pilot test have been compared with the candidates’ scores on IELTS. In addition, VSTEP is developed to cater for Vietnamese citizens’ learning and working needs regarding Vietnamese cultural, economic and social content which is integrated in the test paper...”

The test is composed of four modules - listening, speaking, writing and reading - which are scored separately to produce a composite score corresponding with relevant performance levels of test-takers on the CEFR-VN.

The VSTEP serves a range of general English proficiency purposes in Vietnam, the primary one of which is to screen university graduates on exit. Non-English majors are expected to achieve Level 3 of the CEFR-VN while English majors are assumed to demonstrate higher levels (4 – 5 of the CEFR-VN) (Dunlea et al., 2018). With more than 200 public and private universities in Vietnam and roughly 1.7 million students currently being enrolled, it is expected that the demand for English proficiency certification will increase dramatically, as will the number of test-takers sitting the VSTEP test. This underscores the high stakes of the test which involves various stakeholders and points to the need for validation research. However, there have hitherto been only a few published research articles and book chapters documenting the validation of the

test either at the national or local level (Nguyen, 2020; Nguyen, 2018). Given the stakes of the test and the lack of published research evidence regarding the interpretation and use of its scores, research into any aspects of the validity of the interpretation and use of the test scores is desirable for language testing researchers within and beyond Vietnam. Furthermore, research into the interpretation and use of the VSTEP test scores is a timely response to the call for further validation projects, due to the increasingly widespread use of the test at the national and local levels (Carr et al., 2016). Given the timeline and the resources available, the current project focuses on the interpretation and use of scores obtained on the reading component of the VSTEP test developed at a local tertiary institution (hereafter, university A, or UA) for the purpose of assessing graduates' English proficiency. Therefore, the acronym L-VSTEP reading test is used throughout the thesis.

UA, the research site for the current project, is a large multi-disciplinary university in Vietnam. It offers undergraduate and postgraduate programs to more than 16000 students in 16 faculties with 38 academic disciplines. The English learning curricula at UA were designed in alignment with the six levels of English proficiency as stipulated in the CEFR-VN. The L-VSTEP test at UA was developed by a test design team composed of experienced lecturers and researchers who have taken extensive training courses for test item writers and test raters offered by the national test development team under the auspices of the National Foreign Language Project 2020.

Since 2015, UA has required its students in all disciplines to take an in-house VSTEP (L-VSTEP) test before graduation. English major students are expected to achieve Level 5/C1 while non-English major students, depending on their discipline requirements, should achieve Level 2/A2 to Level 4/B2 in order to meet the English proficiency standards for their graduation. For the purpose of this study, only final-year undergraduate students whose disciplines require English proficiency Levels 3/B1 to 5/C1 and who were preparing to take the L-VSTEP test for their graduation were contacted. The reasons for choosing final-year undergraduates as the target participants for the current project were that they had finished all the language skill modules of their Bachelor programs, and therefore were familiar with the reading text types, genres, and difficulty levels typical of those encountered in the L-VSTEP reading test.

The main theoretical framework adopted for the purpose of validation in this study is the argument-based approach developed by Kane (1992, 1994, 2013) and later used by Chapelle, Enright, and Jamieson (2008) in language testing and assessment. The rationale for choosing this approach is twofold: first, the argument-based approach has been adopted in a number of language

test development and validation projects and proved itself to be simple, flexible and effective (Kane, 2012). The use of this approach is expected to contribute to the building of a coherent argument for the interpretation and use of the L-VSTEP reading test score, with a particular focus on the explanation inference that links the observed scores to the underlying theoretical construct of reading proficiency and the extrapolation inference that links observed scores to the expected scores in the target language use domains. Second, the argument-based validation of the L-VSTEP test is anticipated to provide a unique perspective from the local context of standardized language test validation in Vietnam to the global discussion about the use of this approach in language test validation, which in turn, contributes to the wider use of this line of research inquiry. The next section offers a brief discussion of the key concepts used throughout the study.

1.3. Definitions of key concepts

Validity and validation

This study involves the examination of the interpretation and use of the test scores of a language test. The term *validity* used in the study refers to language assessment contexts and educational measurement more broadly. Validity is a core concept in language testing because for many test users, a valid test means a fair test and the search for a test that has been validated reflects the actual practice of test users and the concerns of all those who are involved with testing practice (Chapelle, 2012a). As discussed later in section 3.2, however, such terms as “valid tests” or “validated tests” are no longer accepted in contemporary approaches to language test validation. There is no single definition of validity that can be used for a variety of tests in a variety of contexts and throughout the history of language testing. Instead, in each of the periods of language testing theory development, a new conceptualization of validity has come into play, bringing opposing views and debate. Validation is defined as the process of justifying the meanings or the interpretations and uses of a testing outcome (Chapelle & Voss, 2013). Just as there exist different conceptions of validity, the process of validation varies across testing contexts and in line with differing conceptualisations of validity. This section discusses some key conceptualisations of validity and the associated validation frameworks during the history of language testing literature.

Validity is defined by such scholars as Cronbach and Meehl (1955b), Lado (1961), Heaton (1975) and Henning (1987) as the extent to which the test measures what it is intended to measure. This definition of validity was used widely at the time and required language test developers and researchers to demonstrate evidence for three types of validity; namely, content validity, criterion-

related validity and construct validity, if the test was to be considered valid. Content validity refers to the expert judgment of the extent to which the test represents the content of tasks within the domain being measured (Fulcher, 1999; Hughes, 1989; Kane, 2013). Criterion-related validity is examined through the comparison of the test under investigation with other tests deemed to measure relevant constructs (Chapelle & Voss, 2013). Two types of criterion-related validity often investigated are concurrent validity and predictive validity. Concurrent validity correlates the test with other measures of similar constructs while predictive validity compares test-takers' performance on the test with their future performance in the target domain (Hughes, 1989; Oller, 1979). Construct validity involves the demonstration that the test is actually measuring the theoretical construct it claims to measure, and as such, the test validation process usually starts with the definition of the construct being assessed (Cronbach & Meehl, 1955a).

The limitations associated with the three-validity approach such as the difficulty in finding proper criterion measures or with defining and operationalizing constructs have driven the need for a more unified framework for investigating validity. As a result, Messick's (1989) unified model of validity came to the fore. Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p.13). Instead of being examined as separate types of validity, content validity, concurrent validity and predictive validity were considered different aspects subsumed under a unitary concept of validity, with construct validity being a central aspect. The validation process is, therefore, the process of accumulating evidence to support the interpretations and uses of test scores on the basis of logical, empirical and ethical considerations (Messick, 1989). Chapelle and Voss (2013) refer to this model as the "evidence-gathering approach" as it places heavy demands on validators who need to take into account multiple theoretical perspectives and build up multiple lines of research evidence for the proposed interpretations and uses of test scores (Kane, 2013). Also, due to its open-ended nature, the approach offers validation researchers little in terms of where to start and when to stop (Kane, 2013).

A more simplified validation process was suggested by Bachman and Palmer (1996) by reframing the conceptualization of validity as "test usefulness" (Chapelle & Voss, 2013). Accordingly, rather than accumulating different types of validity, this approach provided a more practical framework, catering for practitioners, to judge the quality of a test based on six features;

namely, construct validity, reliability, authenticity, interactiveness, impact and practicality. While the approach might be seen as an alternative way of examining validity, it has not been used extensively in language testing literature because the simple alignment of construct validity to “test usefulness” has not kept up with the “mainstream thinking since Messick” (Fulcher & Davidson, 2007).

In order to provide a more coherent and practical framework for language test validators, the argument-based approach to validation was developed through a series of papers by Kane (1992, 1994, 2013), Mislevy, Steinberg, and Almond (2003), Bachman (2005), Bachman and Palmer (2010) and Chapelle et al. (2008). Aspects of validity are no longer investigated exclusively as independent of each others, but instead are subsumed under a two-stage approach: articulating an interpretive argument through a chain of inferences to link the observed performance with the conclusion about a student’s related language proficiency based on that performance; and formalizing a validity argument which evaluates the proposed interpretation and use of test scores. Described as a simpler approach to validation (Chapelle et al., 2008), but still retaining the breadth and rigour of the unified model (Kane, 2013), the argument-based approach has been adopted in numerous language test validation projects (see for example, Aryadoust, 2013; Chung, 2014; Jia, 2013; Jun, 2014; Kadir, 2008; Llosa, 2008; Voss, 2012) and therefore, serves as the central theoretical framework for the current validation project. This theoretical framework will be further developed in Chapter III to underpin the project.

Reading comprehension

Generally, the purpose of validation attempts is to evaluate if the interpretation and use of the test scores are meaningful and are sufficiently informed by relevant empirical evidence and theoretical considerations. As such, in order to examine the validity of the interpretation and use of the L-VSTEP reading test, it is essential to define the reading ability that the test is intended to measure, which is usually referred to as “construct definition” in language testing literature. This section, therefore, offers a brief discussion of reading comprehension and reading ability as the building blocks for the validation of the L-VSTEP reading component.

Basic definitions of reading include “the process of receiving and interpreting information encoded in language form via the medium of print” (Urquhart & Weir, 1998, p.59) and “the ability to draw meanings from printed page and interpret this information appropriately” (Grabe & Stoller, 2013, p.3). However, framing the conceptualisation of L2 reading this way is likely to downplay

the complex nature of reading comprehension and ignore the various component factors and processes that are involved in the reading process. Instead of defining reading in simple terms, Grabe and Stoller (2013) problematized the simple definition of reading by discussing the various issues that need to be addressed if the nature of reading is to be understood. For example, this definition of reading fails to account for the purposes of reading, the various skills, processes, and knowledge bases that are involved in reading, the interface between L1 and L2 proficiency in text comprehension, the contexts in which reading takes place and the various factors that may affect the reading processes such as task features and readers' characteristics. To highlight the complex nature of reading, different perspectives have been called into play, three of which stand out when it comes to the definition of reading construct. These include the processing perspective, the task perspective and the reader purpose perspective (Enright et al., 2000; Grabe & Jiang, 2013).

The processing perspective places emphasis on the conceptualisation of reading abilities in terms of the various linguistic and processing variables employed by readers of differing proficiency levels such as the efficiency in word recognition, working memory, syntactic parsing to name but a few. The task perspective, on the other hand, defines reading construct in terms of task variables that account for performance difference on test items such as the amount of distracting information in the text, the overlap between the wording of texts and items or the concreteness of information in the texts. Related to both task and processing perspectives, the reader purpose perspective conceptualizes the construct of reading according to the purposes of the readers when they engage in a text such as reading for basic information, reading for specific details, and reading for general understanding (Cohen & Upton, 2006). These different perspectives serve as the basis on which theoretical considerations for the validity examination of the L-VSTEP reading are developed, and will be discussed in more details in chapter II.

In consideration of the scarce empirical validity evidence on the interpretation and use of the L-VSTEP reading test, the different approaches to the conceptualization of L2 reading constructs, the need to incorporate these approaches in the context of the L-VSTEP reading test, and the theoretical framework informed by the argument-based approach to language test validation, this study aims to address the following research questions:

1. What reading processes are assumed to correctly answer L-VSTEP reading test items?
To what extent do these processes correspond with the reading processes actually engaged in by test-takers while doing the test?

2. To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with different reading proficiency levels and different academic disciplines?
3. What are the linguistic and discourse characteristics of the L-VSTEP reading texts, items and item-by-text variables? How do these characteristics contribute to the item difficulty of the test?
4. To what extent do students' test scores on the L-VSTEP reading test predict their reading performance in the relevant academic programs?
5. To what extent are reading tasks and skills assessed in the L-VSTEP reading test aligned with reading tasks and skills required in the relevant academic programs?

Development and justification of the research questions will be discussed in Chapter III.

1.4. Significance of the study

Investigation of the research questions formulated above has the potential to make a significant contribution to knowledge. First, since the L-VSTEP test, which is used as evidence of English language proficiency at UA, affects a range of stakeholders, test designers and researchers should be held accountable for demonstrating that the test scores reflect the language ability the test is designed to assess and thus can be meaningfully interpreted and used. Research into the reading component of the test as in the current project enables a theoretically- and empirically informed justification, thus contributing to the meaningful interpretation and use of the test scores. Where the test shows evidence that may weaken the validity of the interpretation and use of the test scores, the validation process has the potential to identify the problematic areas, and remedial actions can be taken as such in the revision and refinement of the test. Second, the current project might offer empirical evidence to enrich the ongoing debates and uncertainties in second language reading comprehension literature such as the notion of L2 reading subskills, the reading processes of the readers and the text and item determiners of L2 reading item difficulty (to be reviewed in Chapter II). Finally, the project draws on the argument-based framework for language test validation to examine the validity of the interpretation and use of the L-VSTEP reading test, thereby contributing to a growing body of research that adopts this framework.

1.5. Thesis structure

Chapter I has introduced the topic, described the context and provided the rationale for the current study. The definition of several key concepts has also been discussed to offer the basis for the development of the study's theoretical foundation in later chapters. Chapter II presents an overview of relevant research literature to guide the development of the theoretical framework for the study in Chapter III and the methods to be employed to address the research questions in Chapter IV. Chapters V through IX present the results and the relevant discussion pertaining to each of the research questions. Chapter X, the validity argument chapter, provides an evaluation of the interpretive argument for the L-VSTEP reading test as laid out in Chapter IV by drawing on the relevant empirical evidence generated throughout the research program. Implications for relevant stakeholders of the test, including test designers, researchers, teachers, students, curriculum designers, and policy makers are also provided in Chapter XI.

CHAPTER II: LITERATURE REVIEW

The primary concern of the current study is to establish validity evidence for the interpretation and use of the L-VSTEP reading test with a particular focus on how the test scores can be attributed to a theoretical construct of English reading proficiency and whether students' scores on the test can predict their self-reported performance in the target language use domains. Therefore, it is essential to consider different approaches to the conceptualization of L2 reading and how these approaches shape the constructs of L2 reading proficiency and inform relevant empirical research in the field. This chapter starts with a discussion of the two general approaches to L2 reading conceptualizations: reading as a process and reading as a product. It then proceeds to explicate how these approaches inform the conceptualization of constructs of L2 reading proficiency by elaborating on three key perspectives: the reading processing perspective which focuses on cognitive aspects of L2 reading in naturally-occurring contexts (ie., non-test context) and the reading task and reading purpose perspectives which focus on L2 reading in assessment contexts (i.e., test context). These interrelated perspectives form an integrated model of L2 reading proficiency (see Figure 2.1), which informs understanding of the construct of L2 reading proficiency assessed in the L-VSTEP reading test (described in section 3.3.1). Empirical evidence pertinent to each perspective is also reviewed.

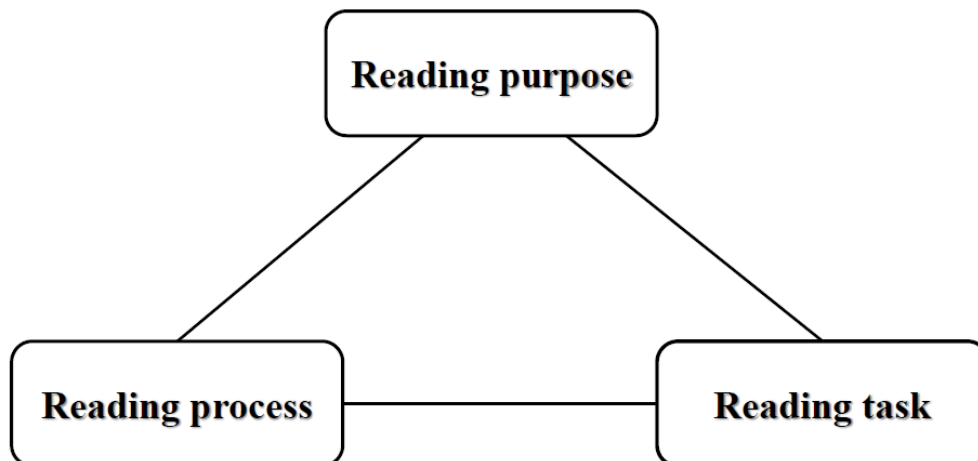


Figure 2. 1. The integrated model of L2 reading

2.1. The nature of L2 reading comprehension

L2 reading is of a complex nature to which contributions from extensive research efforts over the years cannot simply be encapsulated in one or two simple definitions. Different approaches to conceptualizing reading comprehension have been proposed to explain the nature of reading, two of which stand out in the literature, namely reading as a product and reading as a process.

Reading as a product places emphasis on comprehension – the ultimate purpose of reading. This approach entails both quality and quantity of the meaning representation that readers construct as a result of the interaction between their mental activities and the texts (Yamashita, 2002). Conceptualizing reading as a product entails the assumption that readers comprehend a text to the extent that their acquisition of traditional comprehension skills, such as the ability to grasp main ideas, to understand explicit and implicit information, and to make inference are demonstrated by their performance on the text comprehension questions (Myers, 1991). The different explanations above evince a contextualized view of reading in an assessment context where readers' comprehension of a text is measured by their performance on discrete tasks or items related to that text.

The conceptualization of reading as a product, however, is associated with two limitations: the variation in the product and the measures used to assess the product (Alderson, 2000). First, different readers have different interpretations of a text, and these interpretations are influenced by various factors many of which are exclusively independent of the text per se. As stated by Alderson (2000), the text “does not contain meaning which is waiting to be discovered by an able reader” (p.6). Instead, meaning, or what is comprehended by the reader, only comes as a result of the interaction between the readers and the texts to realise “*meaning potential*” (Halliday, 1978; Widdowson, 1979) inherent in the text. Since readers differ greatly in their knowledge, skills and experiences, their interpretation of the same text is likely to be different, even without the influence of other contextual factors. This leads to the second limitation – what measures can be used to assess if a reader has comprehended a text and to what extent this comprehension is considered appropriate. If readers have different interpretation of a text, what criteria can be based on to determine which interpretation is correct and which is not? The perception of “correctness” is problematic here as viewing the interpretation of a text as correct or incorrect may be theoretically misguided (Alderson, 2000). This also poses problems for test constructors because the choice of

specific tasks or items to gauge different levels of comprehension must reflect the conceptualization of what is acceptable and what is not in terms of text interpretations.

In addressing the limitations discussed above, researchers have searched for appropriate methods to systematically link the product of reading – comprehension – with what accounts for that product, and to examine if measures of reading comprehension have been appropriately designed so as to precisely elicit the types of ability or comprehension they are intended to measure. L2 reading research that takes reading as a product primarily involves the examination of the relationship between the test results and the various reader, text and task variables of interest such as the use of correlational and regression analysis in examining linguistic factors that underlie performance consistency or test item difficulty (Freedle & Kostin, 1993; Gorin & Embretson, 2006), the use of factor analysis to uncover the underlying patterns of subskills intended to be measured by the test (Sawaki, Stricker, & Oranje, 2009; Song, 2008), and the use of univariate and multivariate analysis in comparing performance by readers of different ages, genders, proficiency levels, or first language backgrounds (Carrell, 1991; Shiotsu, 2010). A detailed review of these studies is provided in section 2.2.

Unlike reading as a product approach, reading as a process prioritizes the understanding of the processes that readers are engaged in during interaction with a text for the purpose of meaning construction (Yamashita, 2002). Different readers may arrive at the same interpretation of a text through different processes; and conversely, different readers may have different interpretations of a text depending on different processes that they engage in. Myers (1991) reviewed the transaction/interaction models to describe reading as a constructive process during which readers construct meaning through a mixture of information, contexts and the readers' existing knowledge. During this process, several variables come into play, including the readers' purposes, the types and structures of the texts, the context of reading, measures of assessing comprehension and the characteristics of the texts and the readers (Myers, 1991, p.258). While this approach is important for the understanding of the nature of reading, uncovering the underlying processes is not an easy task because reading is perceived as a silent, internal and private process (Alderson, 2000). To tackle this problem, several methods have been employed by researchers of L2 reading, most of which are qualitative in nature. The predominant methodological procedures for this line of research involve introspection and eye-movement.

The examination of eye movement has been used in a number of eye tracking studies (Bax, 2013, 2015; Brunfaut & McCray, 2015a) to investigate the online cognitive processing and strategies of the readers while responding to specific reading comprehension items and to compare the reading processes of readers in test and non-test conditions. On the other hand, introspection techniques such as think-aloud protocols and stimulated verbal recall have been in extensive use for eliciting readers' post-event accounts of their reading process (see Weir, Hawkey, Green, & Devi, 2009 for a review of these studies). Recent methodological and technological innovations in language acquisition research have enabled a combination of both introspection and eye movement methods to triangulate different sources of data for examining readers' reading processes. A detailed discussion of the relevant studies and their associated methodological paradigm is offered in section 2.2.

Although the process-oriented approach has received more attention than the product-oriented approach in the literature on reading comprehension research (Alderson, 2000; Myers, 1991; Yamashita, 2002), researchers commonly believe that both product and process of reading occupy equally important places in the literature, particularly in the field of language testing, due to both practical and methodological considerations. Practically, major international English tests such as TOEFL and IELTS still conform to the reporting practice of separate component skills together with a combined composite score for all modules with their associated written descriptions of the proficiency levels relevant to each score. This practice only documents test-takers' final scores without detailing the specific processes that underlie their performance. As succinctly pointed out by Messick (1989), test-takers' scores can be affected by multiple factors among which test-wiseness and test method effects are irrelevant to the core construct being assessed, which poses threats to the validity of a test. The former refers to the possibility of test-takers' taking advantage of factors peripheral to the test to arrive at a correct answer, while the latter concerns factors associated with the measurement methods rather than the intended ability to be measured by the test. This presses the need for test constructors and validators to take both the scores and processes that induce those scores into account when designing their research. Methodologically, exploring the product and the process of reading independently of each other may only provide a one-sided view into the nature of reading. Combining the product view and the process view, however, can offer a more in-depth insight into the whole enterprise of reading comprehension. It stands to reason that the readers' comprehension of a text should be in

congruence with the processes induced by the cognitive demand of that text. In addition, the combination of quantitative and qualitative techniques within a mixed-method paradigm may provide an essential springboard for the examination of both product and process of reading. For the reasons discussed above, this study considers both approaches: reading as a process and reading as a product through its mixed methods research design.

2.2. Constructs of L2 reading

The current project deals with the concept of L2 reading comprehension in an assessment context, the purpose of which is to examine the validity of a test that measures L2 readers' reading proficiency. Therefore, it is essential to describe what reading proficiency is for a particular purpose and in a particular context before it can be measured. This practice is usually referred to as "construct definition" in second language assessment (Alderson, 2000). Construct is defined as "a psychological concept, which derives from a theory of the ability to be tested" (Alderson, 2000, p.118). Different tests for different purposes and in different contexts require different theoretical definitions of constructs. Previous researchers defined and operationalized the constructs of L2 reading from different perspectives, three of which, namely the processing perspective, the task perspective, and the reader purpose perspective are of particular relevance to the purpose of the current study and are discussed in sections 2.2.1, 2.2.2, and 2.2.3 respectively.

2.2.1. The processing perspective

Informed by the process-oriented approach to the conceptualization of L2 reading comprehension, the processing perspective conceives the construct of L2 reading proficiency as involving different processes and strategies during the reading activity. These processes and the conceptual model that illustrates them are discussed next.

General approaches to describing reading processes

The evolution of research and theories about L2 reading processes over the decades have been influenced by three prominent approaches from L1 reading literature. These approaches, referred to by Grabe (2009) as metaphorical models of reading, include the bottom-up approach, the top-down approach and the interactive approach. The bottom-up approach describes reading as a mechanical process, sequentially starting with the visual analysis of printed materials, followed by the construction of phrases and sentences, and finally, the building of semantic representation of the text (Gough, 1972; LaBerge & Samuels, 1974). This view places particular

emphasis on the processes of word-by-word, letter-by-letter and sentence-by-sentence decoding of text while ignoring the role of prior experience and background knowledge that the reader brings to the text. Failure to comprehend texts is assumed to result from failure to decode visual input. At the heart of the bottom-up approach lie several models of reading including the Verbal Efficiency model (Perfetti, 1985, 1988) and the Word Recognition model (Seidenberg & McClelland, 1989). The Verbal Efficiency model highlights the role of word recognition skill as the main source of individual difference between skilled and less skilled readers. Accordingly, skilled readers have efficient word recognition skill which allows them to free more attentional resources for other cognitive processes during reading comprehension. On the other hand, less skilled readers often have difficulty with higher-level processes due to their inefficient word recognition, spelling problems and weak phonological knowledge (Grabe, 2009). The Word Recognition model, based on connectionist theory which posited that “representations of words are distributed across many simple processing elements” (Snowling, Hulme, & Nation, 1997, p.89), also attaches great importance to word recognition skill in text comprehension. This model views word recognition as a cumulative process of orthographic, phonological, and semantic information that becomes automatised in fluent reading.

Contrary to the bottom-up approach, the top-down approach conceptualizes reading as a reader-driven process during which text comprehension is mainly based on the readers’ conceptual and world knowledge while the role of decoding skills are considered minimal (Goodman, 1986). A prominent model of reading that supports the top-down view is the psycholinguistic guessing game model proposed by Goodman (1967). This model is likened to a hypothesis-testing cycle of generating and confirming expectations, based on the match between the readers’ background knowledge and minimal information retrieved from the texts. As such, breakdown in comprehension is due primarily to the inefficient higher-level knowledge sources, including the ability to elicit semantic and syntactic information, rather than to the lower-level processes of graphophonic decoding.

Neither the bottom-up approach nor the top-down approach gained ascendancy over the interactive approach which adheres to contemporary perspectives about reading processing (Nassaji, 2014). The interactive approach is situated at the interface between the bottom-up and top-down approaches, which underlies the importance of both lower-level processes and higher-level processes in reading comprehension. This approach suggests that the comprehension process

entails at least three subprocesses: a phonological and orthographic decoding process, a semantic and syntactic extracting process, and a text-integration process. Major models of reading aligned with the interactive approach include the Interactive Compensatory Model (Stanovich, 1980, 1984) and the Compensatory Encoding Model (Walczyk, 1995, 2000). The Interactive Compensatory Model proposes that comprehension is contingent upon the combination and integration of both the lower-level and higher-level processes with the incorporation of a compensatory mechanism. In other words, inefficiency at one level of processing can be compensated for by other processes to facilitate the comprehension process. However, the operation of the compensatory mechanism is offset by the limited working memory capacity of the readers as the consumption of cognitive resources for the compensatory activity is likely to result in less resources being left for other comprehension processes (Stanovich, 1982). The Compensatory Encoding Model builds on the Interactive Compensatory Model to introduce time pressure effect as a mediator in the compensatory mechanism. Specifically, under time pressure conditions, lower-level processes become more prominent because higher-level processes are less free to operate (Grabe, 2009).

A large body of research in L1 contexts has unpacked the importance of lower-level and higher-level processes and their associated subcomponents in reading comprehension. Studies with children and adult readers examined and endorsed the crucial role that word recognition plays in the reading comprehension process (Bell & Perfetti, 1994; Cunningham, Stanovich, & Wilson, 1990; Shankweiler, Lundquist, Dreyer, & Dickinson, 1996; Stanovich, West, & Cunningham, 1991) as well as highlighted the role of this skill as distinct from general comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). Unique contribution of higher-level processes to reading comprehension was also documented by research in L1 contexts (Hannon & Daneman, 2001; Landi & Perfetti, 2007; Nation & Snowling, 1998). A third line of research enquiry reported the relative contributions of both lower-level and higher-level processes to reading comprehension among skilled and less skilled readers (Hannon, 2012; Jackson, 2005; Landi, 2010).

While research pertaining to the contribution of lower-level and higher-level processes to reading comprehension has been well documented in L1 contexts, similar studies in L2 contexts are few and far between. A majority of these studies investigated and consistently attested to the critical role of efficient word recognition as major predictors of reading comprehension among L2 young readers (Geva & Wang, 2001; Grant, Gottardo, & Geva, 2011; Verhoeven & van Leeuwe, 2009) and skilled adult readers (Akamatsu, 2003; Nassaji, 2003; Shiotsu, 2010). In

addition, this body of research also found that word recognition can distinguish skilled readers from less skilled readers. Particularly, Nassaji (2003), through discriminant function analysis, found that lower-level process of word recognition including phonological and orthographic skills reliably distinguished skilled from less skilled L2 adult readers, and thus he concluded that lower-level processes were integral components in L2 reading comprehension. Similar findings were also reported by Akamatsu (2005) and Shiotsu (2009) in the Japanese context.

The cognitive processing approach

Building on the bottom-up, top-down and interactive approaches, Urquhart and Weir (1998) introduced a matrix of reading in which two types of reading – careful and expeditious are interwoven with two levels of reading – global and local. This matrix was then adopted and refined by Khalifa and Weir (2009) in their cognitive processing model of reading.

The cognitive processing model (Khalifa & Weir, 2009) illustrates three core components of metacognitive activity, central processing core, and the knowledge base during the reading activity, each comprising various subprocesses. This is illustrated in Figure 2.1.

The metacognitive activity involves goal setting, monitoring and remediation. Depending on the text, the specific requirements of the tasks and the purposes for reading, the reader may decide to engage in careful reading or expeditious reading at either local or global levels. Local comprehension, corresponding with lower-level processes, refers to the understanding of a text at sentence and clause levels while global comprehension, relevant to higher-level processes, indicates the understanding at text-level, across and beyond the micro-structure of clauses and sentences. Careful reading is carried out to extract the complete meanings presented in a text at either the local or global level. This practice is characterized as a slow, careful, linear and incremental process and is taken as the default reading behavior in the building of previous models of reading (Khalifa & Weir, 2009; Weir & Khalifa, 2008). Expeditious reading, by contrast, involves the quick, selective and efficient process of accessing the desired information in a text. Typical techniques of this type includes scanning – reading selectively for extracting specific information in a text, skimming – reading for gist, main ideas or general understanding, and search reading – reading to search for predetermined information at both local and global levels. The integration of these types of reading in the model have acknowledged the valuable contributions of research on the importance of reading speed and reading accuracy (Carver, 1992; Weir et al., 2009) as well as the role that word recognition skills play in reading comprehension as discussed

in the previous section. The comprehension process is constantly monitored in accordance with the reading goals and remediation is triggered whenever comprehension breakdown takes place.

The central processing core details a hierarchical structure of text comprehension from the lower-level processes of word recognition, lexical access and syntactic parsing up to the higher-level processes of propositional meaning establishment, inferencing, mental model building and text-level structure construction. This process depicts the readers' cognitive operation as they engage in a range of activities from matching the visual form with the mental representation of the orthographic form of the words, retrieving lexical entry from the mental lexicon, stringing words together for literal meaning retrieval at clause and sentence levels, to interpreting implicit message inherent in the text, relating adjacent pieces of information, and finally to generating comprehension at discourse level (Khalifa & Weir, 2009).

Two features characterize the hierarchical conceptualization of reading processing. First, informed by the psychological principle of "limited attentional resources" (Perfetti & Lesgold, 1977; Schneider & Shiffrin, 1977), readers' lower-level processes, especially word recognition skill, should become strongly automatized and rapid for a smooth transition to the interpretation at the whole-text level. This is because if readers reserve too much of their attentional capacity for lower-level text decoding, they will have less resources left for higher-level processes (Nassaji, 2014). Automaticity and efficiency are also believed to be the distinction between lower-level and higher-level processes which are more subject to conscious processing (Brunfaut & McCray, 2015b). Second, beyond the textual retrieval of meaning at clause and sentence levels, contextual variables such as prior experience, topical knowledge or knowledge of text structure can be brought to bear on comprehension. These contextual variables may serve to enrich the propositional meaning derived from decoded text or to aid decoding where they are needed (Khalifa & Weir, 2009). These types of knowledge together with lexical and syntactic knowledge are subsumed under the knowledge base component in the model and are at the disposal of the readers to support their comprehension process in alignment with the central processing activities and the goal setting mechanism.

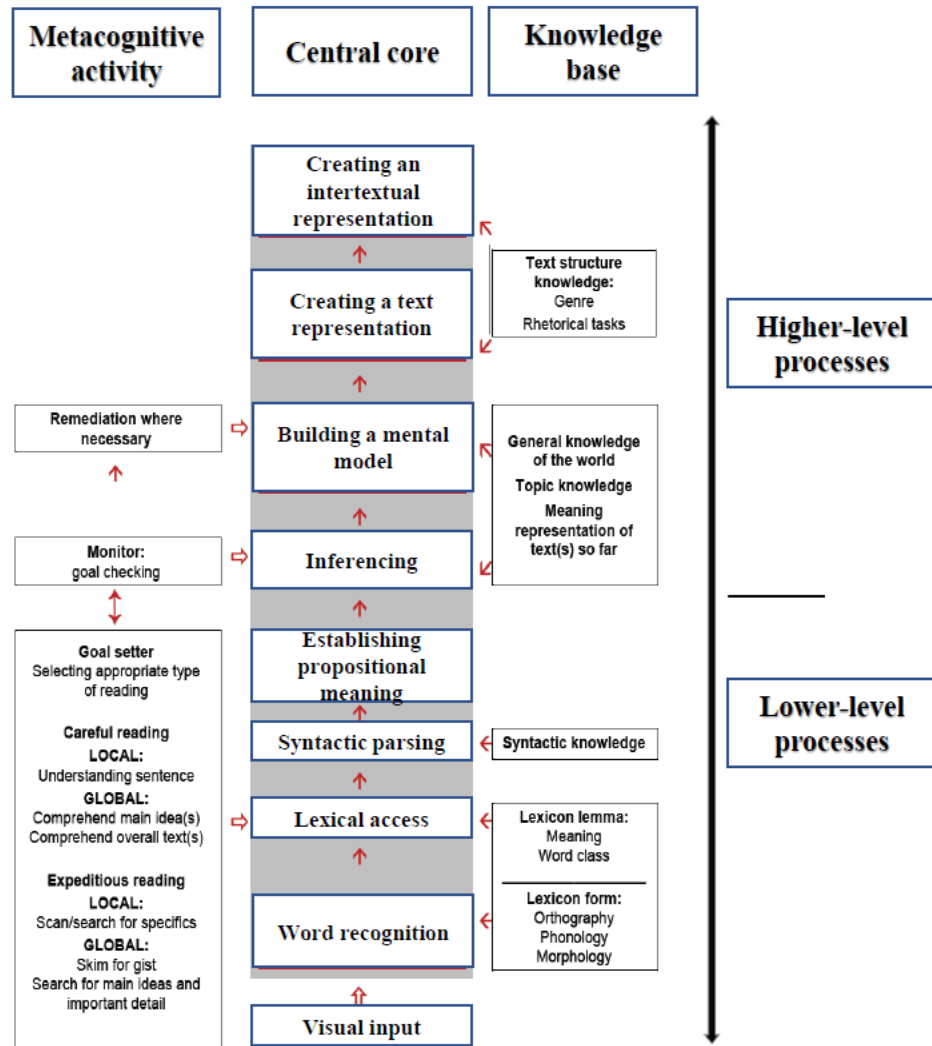


Figure 2. 2. The cognitive processing model (Brunfaut & McCray, 2015b)

In light of the cognitive processing model of reading developed within Weir’s (2005) validation framework, Weir et al. (2009) looked into the congruence between the constructs measured by the IELTS reading test and the actual academic reading practice in the target domain by examining the readers’ cognitive processes. Results of the questionnaire and verbal protocol analysis show that participants applied both expeditious and careful reading types to seek answers to the test items. Furthermore, the responding patterns consistently followed the general approach to academic reading out of the testing conditions in a sequential manner: quick and selective search reading followed by careful reading of relevant texts (Weir et al., 2009). This line of research has paved the way for a growing body of studies that employ eye-tracking technology and verbal report to delve into the cognitive processes of the readers while doing reading tests, which in turn offers implications for the cognitive validity of those tests.

In order to explore the differences in cognitive processing between successful and unsuccessful readers, Bax (2013) interviewed a group of 71 Malaysian students based on traces of their eye movements during an IELTS reading test. Results indicated that proficient and less proficient students differ in terms of the cognitive processing at the lexical and grammatical levels. However, there was no evident differences recorded at the higher-level cognitive processes. This is, as explained by the researcher, in part due to a lack of specific items targeting at this level. Another finding of the study was the superiority of eye-tracking technology as a methodological innovation to differentiate between successful students who were able to read selectively and locate quickly particular areas of a text for specific information and unsuccessful students who were not able to do so. In a similar study with a different cohort of multinational participants (N = 41), Bax (2015) confirmed findings from the previous study. He also found that those students who were not successful with items of higher-level processing failed to locate areas for specific information and so failed to make correct inferences. Similar findings from the two studies suggested that the reading pattern discovered in the earlier study was not confined to participants of a particular nationality but rather, applied to participants of different nationalities and different languages. This underscored the value of replication research in language testing.

In a similar attempt to combine both eye-tracking technology and stimulated verbal recall, Brunfaut and McCray (2015b) unpacked the cognitive processes of test-takers while taking the APTIS reading test with respect to tasks and participants of different proficiency levels. They identified a range of cognitive processes at both lower and higher levels noted in Khalifa and Weir (2009) model with the exception of intertextual representation. Item completion was also associated with careful and expeditious reading types at either global or local levels depending on task types. They concluded that the APTIS reading test measured the constructs relevant to the reading behaviors in the target domain in terms of the cognitive processing. A similar study was conducted by Bax and Weir (2012) on the CAE reading test and generated similar findings supportive of the cognitive validity of the test.

The discussion of research findings from various studies above highlights several noteworthy points. First, the cognitive processing model (Khalifa & Weir, 2009) proved to be a useful framework for investigating the types of reading and the degree of cognitive engagement of readers at different proficiency levels, which offers promising premises for language test validators. As argued by Alderson (2000), “the validity of a test relates to the interpretation of the

correct responses to items, so what matters is not what the test constructors believe an item to be testing, but which responses are considered correct, and what processes underlie them” (p.97); this model, therefore, provides straightforward guidelines to uncover these underlying processes. Second, the use of eye-tracking technology has dominated this line of research methodologically. While this modern technology is capable of providing accurate accounts of readers’ eye movements while reading, it is not without its limitations (See chapter IV for a detailed discussion of these limitations). The last noteworthy point related to the studies reviewed above is that readers’ cognitive processes should not be discussed in isolation from the cognitive demands and characteristics of the relevant reading tasks because the types of tasks determine the level of cognitive engagement of the readers. This is dealt with in the next section.

2.2.2. The task perspective

Of particular relevance to L2 reading assessment contexts is the task-based approach to conceptualizing reading ability. Alderson (2000) argued that in order to know if a reader has comprehended a text and how well he or she has comprehended that text, it is necessary to task the reader with some sort of reading activity and elicit aspects of his or her reading performance in some way. One approach is to relate readers’ reading ability in the test conditions with their actual reading ability in real world on the basis of their performance on a reading task. Carroll (1993) defined a task as “any activity in which a person engages, given an appropriate setting, in order to achieve a specifiable class of objectives” (p.8). Later, Alderson (2000) adapted this definition for second language assessment contexts and used the term *language use task* which refers to “an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular setting”. Based on Bachman’s (1990) test method facets framework, Alderson (2000) proposed five essential aspects of a task including characteristics of the setting, characteristics of the test rubric, characteristics of the input, characteristics of the expected response and relationship between input and response, among which input, expected response and their interaction are of particular importance to, and have been an avenue for, much research effort in reading assessment. As argued by Enright et al. (2000) in their search for an optimum framework for the TOFEL reading test, “it is useful to develop a set of text and task variables that can account for the variance in performance that occurs on reading test questions” (p.3). These variables may serve as an interpretable description of the factors that bring difficulty for readers and direct them to remedial actions needed for successful performance on reading tests.

These text and task variables and research that has been done on them are discussed in the following sections.

Text variables

L2 reading comprehension occurs as a result of the readers' interaction with the texts and, therefore, is partly influenced by the characteristics of those texts. Alderson (2000) reviewed a number of text variables that affect the reading process, several of which are of particular salience to the reading assessment contexts in the current study. These are text topic and content, text type and genre, and text readability.

Text topic and content

It is commonly assumed that readers' comprehension of a text is influenced by how much they know about the topic and content of the text. While several theories, such as the Schema theory and the Construction-Instruction theory (Nassaji, 2002), have been proposed to account for the readers' variable of background knowledge brought to the text comprehension, no such theories exist to elucidate the difficulty of texts as a feature of their content and topic (Alderson, 2000). Instead, researchers have attempted to examine the levels of abstractness and the topic familiarity of texts and relate these to how readers fare on relevant reading tasks. It was generally found that concrete texts tended to be more comprehensible, interesting and memorable than abstract texts (Sadoski, Goetz, & Fritz, 1993a, 1993b; Sadoski, Goetz, & Rodriguez, 2000; Sadoski & Paivio, 2001) and that familiarity with the content and topic of texts aided reading comprehension (Alptekin, 2006; Alptekin & Erçetin, 2011; Horiba & Fukaya, 2015) although this effect was found to vary with levels of comprehension and types of text. For example, Alptekin (2006) found that the localized version of an American short story facilitated Turkish L2 readers' inferential comprehension but not their literal understanding. Similarly, Alptekin and Erçetin (2011) reported that textually and contextually localized texts improved inferential comprehension but did not affect literal understanding in L2 reading.

Text type

It is often noted that comprehension difficulty may arise not because of the actual content of the text per se, but rather from the way the text is written and the various features that make one text different from another (Alderson, 2000). Text type, therefore, is usually considered a major predictor of variance in reading comprehension. Although several text genres have been documented in the literature such as recount, persuasive texts, transactional texts, narrative,

process descriptions and expository texts (Barrot, 2016; Frønes, Narvhus, & Aasebø, 2013) a substantial body of research has been devoted to the differentiation between narrative and expository texts and the relative contribution these two types make to reading comprehension. Research often found favorable effects of narrative texts over expository texts on reading comprehension (Alderson, 2000; Saadatnia, Ketabi, & Tavakoli, 2017). Explanatory variables for the difficulty associated with these two text types were also identified. McCormick and Zutell (2007) attributed the difficulty of expository texts to text structure, technical vocabulary, abstract concepts, novel proposition and readability of texts. Gardner (2004) noted that narrative texts were based on familiar topics on which readers may possess some background knowledge while expository texts may appeal to only a subset of readers. Gardner (2004) also discovered that expository texts contain more low-frequency, topic-specific vocabulary than narrative texts, which is likely to impose a heavier cognitive load on readers' cognitive processes, especially those at the text decoding levels. Best, Floyd, and McNamara (2008) reported that word recognition was the most powerful predictor of narrative text comprehension while background knowledge contributed the most to comprehension of expository texts.

Text readability

Innovations in corpus linguistics and computational linguistics over the years have driven the development of various formulae, indices and other procedures to account for the numerous linguistic variables that affect text difficulty (Alderson, 2000). These indices provide quantitative account of such features as lexical characteristics, syntactic complexity, and coherence and cohesion of texts (Barkaoui, 2015). Lexical aspects are usually quantified in terms of lexical density (the proportion of content words in a text), lexical variation (the ratio of the total number of different words and the total number of words in a text, usually referred to as Type-Token Ratio) (Laufer & Nation, 1995), lexical sophistication (the ratio of low-frequency and high-frequency words) (Laufer & Nation, 1995), and word information such as word frequency, word familiarity and word polysemy (McNamara, Crossley, & McCarthy, 2010). Syntactic complexity concerns the proportion of the amount of information and the grammatical units that surface in a text (Barkaoui, 2015; Lu, 2011). Common measures of syntactic complexity include sentence length (average number of words per sentence) and syntactic similarity (the consistency of syntactic forms across sentences in a text) (Graesser, McNamara, Louwerse, & Cai, 2004). Cohesion was defined by Graesser et al. (2004) as the use of explicit features, words, phrases and sentences in

connecting and interpreting ideas in a text. Two measures of textual coherence are often used: inferential cohesion (level of co-reference among words in a text) and conceptual cohesion (how similar different concepts in a text are) (Green, Ünalı, & Weir, 2010).

Task variables

Another line of research enquiry into factors causing difficulty for readers focuses on the test methods that are used to gauge readers' reading performance. A number of test methods or techniques were reviewed by Alderson (2000), among which multiple choice questions (MCQ) have been considered to be by far the most common way to assess reading (p. 204). This section, therefore, offers a brief review of research on the test method effects on reading comprehension performance, particularly the use of MCQ in L2 assessment contexts, which is relevant to the purpose of the current study.

Test method effects

In order to assess readers' performance on a reading comprehension task, a number of testing techniques have been proposed including, the use of MCQ, cloze tests, gap-filling tests, matching, and ordering. Each of these techniques assesses a specific skill or knowledge area related to comprehension of the text at hand, and by virtue of this, may engage readers in different cognitive processes. No matter what technique is used, its effect on readers' performance must be limited to a minimum level because what is of concern is the readers' reading ability rather than their ability to deal with specific testing techniques. As stated by Bachman (1990, p.111) " if we are to develop language tests appropriately, for the purposes for which they are intended, we must base them on clear definitions of both the abilities we wish to measure and the means by which we observe and measure these abilities". Test method effects, therefore, occur when a test, employing different methods to measure an intended ability, yields different results. In other words, students' scores on the test "are more the result of test methods than of the trait being measured" (Shohamy, 1984, p.147). Examining the nature and reducing the effects of testing methods used in reading comprehension assessment is, therefore, a major concern for language test designers and researchers. Researchers have primarily been concerned about three areas: how different methods affect reading comprehension performance, how a specific test method influences reading processes in test and non-test conditions, and what predicts item difficulty: features of test methods or features of the texts?

Studies on differential test method effects

L2 studies that compare the differential effects of test methods on reading comprehension performance thus far have produced mixed results. Shohamy (1984) compared the effects of MC and open-ended questions presented in either L1 or L2 on L2 reading comprehension by EFL readers and found that MC questions were consistently easier than open-ended questions. This effect was found to be more noticeable in less proficient readers. Kobayashi (2002) compared EFL Japanese readers' performance on different test methods: cloze tests, open-ended questions and summary writing with regards to different text types and different proficiency levels. She found that well-structured texts induced the highest performance on summary writing, lowest performance on cloze tests and made no difference on open-ended questions. In addition, high-proficiency readers were more subject to the influence of different test methods. She concluded that different test methods measured different aspects of reading comprehension. With a similar focus on test method effect but on different cohorts of readers, Zheng, Cheng, and Klinger (2007) investigated ESL/ELD (English as a second language/English literacy development) and non-ESL/ELD students' reading performance on three different question types: MC questions, constructed-responded questions (CR) and constructed-response questions with explanation (CRE). They discovered that while non-ESL/ELD students performed better on all three question types, the difficulty pattern was similar in both groups, participants scored the highest in MC questions, lower in CR questions and lowest in CRE. In a more comprehensive attempt, In'nami and Koizumi (2009) conducted a meta-analysis on the effects of MC and open-ended questions on both L1 and L2 reading. The results suggested that MC questions were easier than open-ended questions in L1 reading. However, no significant differences were found in L2 reading.

A commonality among the studies reviewed above is that test method does affect reading comprehension with the moderation effects of various factors such as readers' L2 proficiency, texts types and L1 backgrounds. Even different types of question (literal/inferential) within a method activate different comprehension and response processes (Kobayashi, 2002; Rupp, Ferne, & Choi, 2006; Sheehan & Ginther, 2001). It is, therefore, important to not only discover how many responses are correct but also what processes underlie those responses and whether these processes reflect the skills the test items are aiming to test.

Studies on test method effects on reading processes

Studies on readers' responses to different test methods generally found additional processes beside those required of the texts or intended by test designers. Research has focused on multiple

choice questions which have been widely used in reading tests and which raise the greatest concern pertaining to validity (Farr, Pritchard, & Smitten, 1990).

Cohen and Upton (2007) compared strategies used by readers to respond to the new TOEFL reading test and traditional multiple-choice tests. Results of the verbal report analysis showed that while some test-takers did actually employ academic reading skills for both local and global understanding of texts as intended by the test designers, others engaged in a range of test-taking strategies to complete the tasks successfully rather than to learn, use or gain anything from the texts. Fortunately, these test-taking strategies were of test-management rather than test-wiseness nature, which did not seriously impact the construct validity of the test. Rupp et al. (2006) investigated, through think-aloud protocols and elicited feedback, the strategies and skills readers employed in responding to multiple choice questions and how these strategies and skills were influenced by characteristics of the texts and questions. They found that readers approached multiple choice questions as a problem-solving task rather than a comprehension task and that they utilized a range of strategies and skills, particularly at microstructure representation level of the texts, to select the answers. They concluded that neither general models of reading comprehension processing nor models for responding to multiple choice questions accounted for test method effects and suggested the development of separate models of response processes relevant to specific question types within a test method.

In summary, studies on item response processes did acknowledge the effects of test methods on the range of strategies and skills that readers employed. These findings provided evidence for a better understanding of readers' behaviors while reading and whether these behaviors match their responses to comprehension items, as Farr et al. (1990, p.211) cautioned "some readers pick correct answers for wrong reasons and wrong answers for correct reasons". Yet, limitations did exist regarding the qualitative nature of the methods used to elicit data. Cohen and Upton (2007) cautioned that task performance may be distorted by reactive effects of verbal report and by readers' more conscientious efforts than in normal conditions. Therefore, instead of taking readers as the main source of data elicitation, other researchers examined what influenced readers' response processes by looking at the interaction between item completion and features of texts and items through the lens of cognitive processing perspectives.

Studies on item difficulty modelling

Embretson and Wetzel (1987) proposed a cognitive processing model for the multiple choice paragraph comprehension item through the quantification of sources of cognitive complexity. The model entails two processing stages: a text representation process where text comprehension occurs and a decision process where answer to an item is selected. The processing stage consists of two activities: lexical encoding and coherence processing whose difficulty can be respectively measured by word familiarity and propositional density. The decision stage is composed of three events: lexical encoding and coherence processing of the alternatives, which is similar to that of the processing of the text; evaluating the alternatives by matching them with relevant information in the text; and justifying the truth status of the alternatives through the processes of falsification and confirmation (Embretson & Wetzel, 1987, p.178). Variables that affect the decision stage may include word familiarity and propositional density of the stems and alternatives, and proposition overlap between the stems, alternatives and the text. The falsification and confirmation processes are perceived to constitute the strongest predictors of item difficulty in the model because correct options that are directly confirmed or incorrect options that are explicitly contradicted by the information in the texts require little processing and thus are easier (Gorin & Embretson, 2006) than those which require inferential information or comprehension at text levels. Researchers operationalized this model by quantifying the variables underlying the two component stages and determined the extent to which they contribute to item difficulty through item difficulty modelling.

The above model informed a number of studies on reading item difficulty modelling. Ozuru, Rowe, O'Reilly, and McNamara (2008) investigated the contribution of item and text characteristics to item difficulty on the Gates-MacGinitie reading tests (GMRT) for 7th – 9th grade and 10th – 12th grade students. They found that text features significantly accounted for item difficulty on the test for 7th – 9th grade student group but not for the 10th – 12th grade student group. Freedle and Kostin (1993) coded 12 item, text and item-by-text variables to examine the extent to which these variables explained the difficulty of 213 multiple choice items taken from 100 TOEFL reading passages. Results suggested that text and item-by-text variables accounted for a relative large amount of variance in item difficulty, thereby providing evidence to support the construct validity of the test.

The aforementioned finding was not supported by Gorin and Embretson (2006) who employed a complicated coding scheme at items, texts and item-text correspondence levels to

identify the features that accounted for processing difficulty of the Graduate Record Examination – Verbal (GRE-V). They found that item difficulty of the test was primarily explained by the decision processes variables necessary for matching information from the response alternatives with that of the texts while no significant differences were detected for the text-processing variables. This suggested a mismatch between the construct defined by the test designers and the empirically derived one because the items failed to capture the range of comprehension skills related directly to the texts.

Informed by corpus linguistics and computational linguistics, Barkaoui (2015) employed both text-analysis software and expert judgment to examine the linguistic and discourse characteristics of the Michigan English Test reading texts and items and how they explained the item difficulty. Twenty two item, text and item-by-text variables were coded and subjected to multilevel modelling analysis. Results shown that four text variables (text length, connectives density, section and non-verbal information) and five item variables (question word familiarity, item reference, subskills tested, explicitness of information requested, and number of plausible distractors) contributed significantly to the item difficulty estimates. However, several construct-irrelevant variables such as item length, item vocabulary level, and degree of lexical overlap were found to be significantly associated with item difficulty while other construct-relevant variables at text levels including syntactic complexity, lexical characteristics, coherence and cohesion, text concreteness and text readability contributed minimally to item difficulty. These findings offered important implications for test design improvement. For example, effects of construct irrelevant factors found in the study can be mitigated or eliminated in future tests design by the standardization of those text and item features across test forms.

To summarize, the studies above provided mixed results regarding the relative contribution of text and item variables to reading comprehension item difficulty. Yet, they generated valuable information for the construct validation of reading tests in L2 contexts and guided test item writers towards timely remedial actions where construct irrelevant factors impacted on test scores.

2.2.3. The reader purpose perspectives

Another important facet of reading comprehension that should be taken into account is the reader's purpose. A reader approaches a reading text with different purposes in mind, which in turn determine the way s/he reads it, the skills s/he uses, the cognitive processes s/he engages in and the ultimate understanding and recall of the text (Alderson, 2000; Enright et al., 2000; Rupp

et al., 2006). The processing perspective and the task perspective discussed earlier can, therefore, be directly related to the reader purpose perspective. This section discusses the different purposes for reading as mentioned in the literature, how these purposes take shape in reading test, and the notion of reading subskills as a corollary of the reader purpose perspective.

L2 reading purposes

There are no exhaustive lists of reading purposes available in the literature as different scholars take different approaches to conceptualizing reading purposes (Alderson, 2000; Grabe, 2009; Linderholm & van den Broek, 2002; Urquhart & Weir, 1998). Of these, Grabe (2009) proposed a well-known list of academic purposes for reading which includes six aspects: reading to search for information, reading for quick understanding, reading to learn, reading to integrate information, reading to evaluate, critique and use information, and reading for general comprehension. Reading to search involves skimming through the text for possible locations of desired information and scanning to find the suitable details. Reading for quick understanding primarily requires readers to use skimming to quickly grasp the gist of a text or to identify areas of the text that need more focus. This usually happens when the reader is under time pressure and quick decisions need to be made regarding a particular level of comprehension. Reading to learn is more appropriately situated in an academic setting. This purpose of reading is perceived as placing more processing demand on the readers as they need to not only understand the literal meaning of the text, but also conceptually organize the content in alignment with what is presented in the text and are able to recall important information for task completion or future use. Reading to integrate involves the synthesis of information across multiple texts or multiple parts of a text. It is more challenging than the previous types as readers need to read selectively and evaluate information retrieved from the texts by incorporating their prior knowledge or experience. Reading to evaluate, critique, and use information takes a step further to require readers to evaluate and critique important or controversial information from the texts. Apart from text comprehension and prior knowledge, the readers' emotional state, interest, attitude and preferences are also brought to bear on completion of tasks that may involve other output activities.

While different purposes for reading can be proposed and classified based on theoretical and practical considerations in language learning and teaching, questions remain as to how to operationalize the constructs of reading for assessment if they are driven by a reader purpose perspective. Among the major international language proficiency tests, the TOEFL test adopts as

its core principles for reading item construction a purpose-driven approach. In building a conceptual framework for the TOEFL 2000 reading module, Enright et al. (2000) proposed four types of reading, or “purposes for reading” for item construction. These include reading to find information (search reading), reading for basic comprehension, reading to learn, and reading to integrate information across multiple texts. They argued that these purposes can be directly related to the task perspective and processing perspective, thus forming a coherent interpretation of the constructs for item building. For example, the *reading to find information* items require the lower-level processes of rapid, automatic recognition of words, working memory efficiency and reading fluency in terms of the cognitive demand, while relevant item types may consist of searching for and matching discrete information or basic details. Likewise, the *reading to integrate information* type imposes an increasingly high demand on readers’ processing as they need to draw on their long term memory, conceptual representation and theories of learning to build an intertextual model of comprehension. This necessitates such item types as integrating information or generating conceptual organization of ideas.

It is conceivable from the discussion above that each purpose for reading demands different combinations of cognitive processes and different item types for comprehension assessment. An inevitable question arises from this: whether reading can be considered a unitary, single notion of overall reading ability which can be measured by different item types or whether reading is a divisible concept consisting of various subcomponents or subskills, each can be measured by separate item types? It is this question to which the following section turns.

Reading subskills

One of the most important aspects of second language reading comprehension is the notion of subskills. Various taxonomies and lists of reading skills have been proposed based on theoretical and practical considerations, with or without empirical justification (Buck, Tatsuoka, & Kostin, 1997; Carroll, 1980; Davis, 1968; Munby, 1978). Despite the fact that the very subskills proposed vary widely across lists and researchers, and that criticism has been raised regarding the lack of empirical evidence or unclear description and distinction of skills (Alderson, 2000), it has become a common practice for language teachers, practitioners and language test designers to use these sets of reading skills or components as the guiding frameworks for course syllabus design, instructional materials development and test item construction (Alderson, 2000; Alderson & Lukmani, 1989a; Lumley, 1993; Song, 2008). Alderson (2000), however, cautioned that the

primary concerns of researchers should not be about how many skills can be listed, but how they can be identified and classified in reading tests. As yet no consensus has been reached as to how reading skills can be categorized or even if separable skills actually exist at all (Alderson, 2000; Rupp, 2012; Tengberg, 2018). Two opposing views have muddled this line of research inquiry: reading is a unitary concept and reading is a divisible concept.

The unitary view

In L2 reading, Schedl, Gordon, Carey, and Tang (1995) investigated whether the reasoning (higher-level) items of the TOEFL reading test measured unique traits beyond linguistic and general discourse competence. They found that the distinction between reasoning and nonreasoning items was not supported as all item types contributed equally to the overall measurement, thus attesting to the unidimensionality of the TOEFL reading test. Sawaki et al. (2009) examined the dimensionality of the TOEFL iBT reading test by considering three traits (basic comprehension, reading to learn, and inferencing) and three methods (three item sets associated with three reading passages) in a Multitrait-Multimethod analysis of four models: A correlated trait/correlated method model, a correlated trait/uncorrelated method model; a correlated trait model; and a correlated trait/correlated uniqueness model. Again, the single trait model was chosen as the final solution, suggesting the unidimensionality of the TOEFL iBT reading section. In a recent project, Tengberg (2018) asked eleven Swedish teachers to classify items of a national reading test in Sweden according to four reading categories: retrieve explicitly stated information; make straightforward inferences; integrate and interpret information and reflect; and examine and evaluate content, language and textual elements. Little consistency was found among the teachers in terms of item classification, which suggested that classifying items according to reading processes/subskills might be a difficult task for teachers and needed more empirical justifications.

The multi-trait view

The multi-trait view of reading comprehension was also supported by numerous studies in the L2 reading research literature. Lumley (1993) asked a group of five experienced EAP teachers to match nine proposed subskills to 22 items of an EAP reading test and judge the difficulty of those subskills and items. The findings demonstrated that there was not only a high level of agreement among teachers about the nine subskills measured by the test items, but also a significant correlation between teachers' perception of the difficulty of the subskills and the item

difficulty based on the Item Response Theory analysis. Song (2008) attempted to explore the divisibility of comprehension measured in L2 reading and listening sections of the WB-ESLPE (Web-based English as a Second Language Placement Exam). Accordingly, she proposed and tested a series of unitary (general reading ability), two-subskill (understanding explicitly stated information and understanding implicitly stated information) and three-subskill (understanding main/topical ideas, understanding supporting/specific details, making inference) models, using structural equation modelling. Results indicated that the two-subskill model fitted data better than the three-subskill model for the reading section while the three-subskill model achieved better fit than the two-subskill model for the listening section. This was indicative of comprehension divisibility being more subject to listening measures than reading measures. Kim (2009) proposed and empirically tested three models of L2 reading subskills, using data from 298 ESL learners' performance on the CEP (Community English Program) reading test at Teacher College – Columbia University: a unitary model, a two subskills (reading for literal meaning and reading for implied meaning) model, and a three-subskill (reading for literal meaning, reading for implied meaning with endophoric reference, and reading for implied meaning with exophoric reference) model. Results indicated that while Exploratory Factor Analysis suggested a unitary concept of reading, Confirmatory Factor Analysis indicated that all three models fitted data well. Although a three-subskill model was chosen as the final solution, no clear reasoning was provided, thus rendering the results open to different interpretations.

Findings from the studies reviewed above underscore the fact that the dimensionality of L2 reading tests remains a highly controversial issue. To find a definitive answer to the question of whether or not L2 reading is a divisible concept as implied by test scores requires extensive research and empirical evidence from different contexts and with learners of different educational, first language and English proficiency backgrounds. This is partly addressed in the current study in the Vietnamese EFL context because one of its aims is to uncover the underlying pattern of the L-VSTEP reading test scores to see how it is aligned with the theoretical construct of L2 reading proficiency as informed by the test development guidelines. In Chapter III, the major theoretical frameworks that guide the exploration of the research problems will be established by the integration of the three prominent L2 reading perspectives discussed above, and on the basis of current and relevant conceptualizations of second language test validation.

2.3. Chapter summary

This chapter has discussed the empirical, conceptual, and theoretical issues related to second language reading comprehension. Three prominent approaches to the conceptualization of the construct of L2 reading and the empirical studies that support them have been reviewed. The processing perspective informs studies that explore the cognitive processes of the readers while they engage in a reading activity. The desire to tap into readers' cognitive processes necessitates the use of introspective methods, a majority of which are subject to inevitable limitations. The task perspective draws attention to the task and text features that may affect the reading outcome, either in test or non-test contexts. This implies that L2 reading assessment research should not only focus on how difficult or easy a reading test is, but what makes it difficult or easy and how to identify the factors that educe the difficulty of the test, as well as whether students' performance on the test given specific task features is comparable to their performance in the non-test context. The reader purpose perspective highlights the controversial issue of whether L2 reading is a unitary or multi-trait concept and whether this concept is exhibited in a reading test. These different perspectives present research gaps that can be addressed in the current project via the formulation and investigation of the five research questions:

1. What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes correspond with the reading processes actually engaged in by test-takers while doing the test?

2. To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with differing reading proficiency levels?

3. What are the linguistic and discourse characteristics of the texts, items and item-by-text variables of the L-VSTEP reading test? To what extent do they contribute to item difficulty?

4. To what extent students' scores on the L-VSTEP reading test predict their reading performance in the relevant academic programs?

5. To what extent are the reading tasks and skills assessed in the L-VSTEP reading test aligned with the reading tasks and skills required in the relevant academic programs?

While the integrated model of L2 reading comprehension as reviewed in this chapter informs the understanding of L2 reading proficiency as measured in the L-VSTEP reading test, there exists the need for an integrative framework underlined by assessment theories to enable a

systematic manipulation of methods that delve into the construct of L2 reading proficiency as measured in the L-VSTEP reading test as well as the extent to which this construct is realized through students' performance in the test and non-test contexts. One such overarching framework that can inform the current study is the argument-based framework proposed by Kane (2013) who drew on Toulmin's (2003) argument structure to lay out a blueprint for the examination of the validity of the interpretation and use of measurement instruments. The framework was then adapted by numerous L2 scholars to examine the validity of the interpretation and use of test scores of various second language proficiency tests. For example, the argument-based framework has been used in validation studies of L2 writing skills (Becker, 2018; Johnson & Riazi, 2017; Lallmamode, Daud, & Kassim, 2016; Ranalli, Link, & Chukharev-Hudilainen, 2017; Yan & Staples, 2020), L2 speaking skills (Brooks & Swain, 2014; Farnsworth, 2013; Huang, 2016; LaFlair & Staples, 2017), and L2 listening skills (Aryadoust, 2013; Li, 2013; Pardo-Ballester, 2010). Few studies, however, have extended the application of this framework to the validation studies of L2 reading tests. The current study contributes to the L2 reading assessment literature in this regard by drawing on the argument-based framework to articulate and empirically evaluate the proposed interpretation and use of the L-VSTEP reading test scores. A detailed discussion of the framework as well as the proposed interpretation and use of the test scores of the L-VSTEP reading test based on the framework is presented in the next chapter.

CHAPTER III: THEORETICAL FRAMEWORK

This chapter delineates the theoretical framework that underpins the proposal and evaluation of the interpretation and use of the L-VSTEP reading test scores. In consideration of a suitable theoretical structure that underpins this validation research, the argument-based approach (Chapelle et al., 2008; Kane, 1992, 1994, 2012, 2013; Mislevy et al., 2003) is chosen for two reasons. First, the argument-based approach, which adheres to the contemporary conceptualization of language test validation, offers a coherent framework for the articulation and evaluation of the validity arguments for the interpretation and use of a language test. It has proved to be a practical, pragmatic, and useful framework for language test validation (Aryadoust, 2013; Brennan, 2013), and is characterized as a simpler and more specific conceptual infrastructure for examining test score interpretation and use (Chapelle, 2012a) than the unitary model of validity (Messick, 1989). Second, numerous language test development and validation projects have successfully adopted the argument-based approach (see for example, Aryadoust, 2013; Chung, 2014; Jia, 2013; Jun, 2014; Kadir, 2008; Llosa, 2008; Voss, 2012), thus not only widening this line of research inquiry at a global scale but also offering a potentially rich methodological springboard for the current project.

The argument-based framework is built upon the conceptual principles of the practical argument structure proposed by Toulmin (1958). Therefore, the chapter begins with a brief discussion of Toulmin's model, followed by the description of the argument-based approach, with a particular focus on the explanation inference and the extrapolation inference – the essence of the study. The chapter concludes with an articulation of the interpretive argument for the L-VSTEP reading component, which serves as the guiding principles for the evaluation of the validity of the interpretation and use of the test scores.

3.1. Toulmin's argument structure

As discussed earlier, the field of language testing has been characterized by the evolution of different conceptualizations of validity and the associated collection of cumulative evidence in the validation process. The criterion model, the content model, and the construct model of language test validity have respectively gained ascendancy and been subsumed by later models, which culminated in the unitary model of validity by Messick (1989). Central to the unitary model is

construct validity which is supported by the accumulation of various types of validity including content validity and criterion-related validity (e.g, concurrent and predictive validity). This model, while providing a theoretically-sound and conceptually-reasonable framework for examining validity, does not indicate where to start the validation process, how to go about collecting evidence for each type of validity, and when to stop the process (Kane, 2013). In other words, practical specifications had not been described in enough details so as to guide language test designers and validators in the development and validation process. This requires a more straightforward, specific, and transparent framework for the examination of language test validity germane to the interpretation and use of test scores. In this connection, Toulmin’s (1958) description of presumptive reasoning (Kane, 2013) or formal/practical arguments (Chapelle et al., 2008) was considered a relevant and viable starting point for the building of a new framework for language test validation. Toulmin (1958) proposed a model of argument structure including six components: Data, claim, warrants, qualifiers, rebuttals, and backing. These components are often laid out as linked to each other via logical connectors (e.g. so, unless, since) within a diagrammatic structure, which preserves the logical reasoning flows from one part to another. Figure 3.1 illustrates a typical Toulmin’s argument structure.

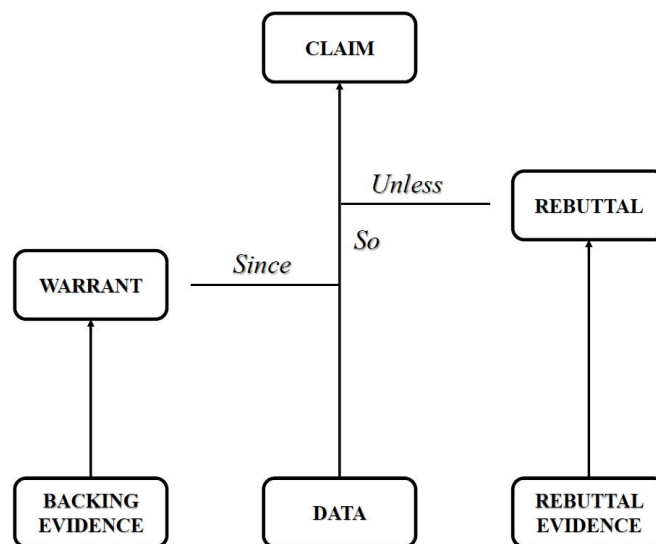


Figure 3. 1. Toulmin’s Argument Structure

A claim is the proposition or conclusion that one wants to make and whose merits one is seeking to establish (Toulmin, 2003). The foundation upon which a claim is made is the data which constitute facts or evidence through investigations (Toulmin, 2003). The arrow from the data to

the claim represents an inference which is authorized by a warrant. Warrants are general, hypothetical statements that legitimize the inference from the data to the claim. In other words, warrants serve as a logical bridge connecting the data to the claim. Depending on the nature of the claim, there may be more than one warrant or inference, each may entails further assumptions. Warrants are supported by backing which consists of “other assurances, without which the warrants themselves would possess neither authority nor currency” (Toulmin, 2003, p.96). Backing may take the form of theories, research, data or experience (Mislevy et al., 2003) and may vary with respect to the field of argument and the elaboration of warrants (Toulmin, 2003). Whereas some warrants are highly tenable a priori and do not require much backing, most warrants need backing to establish their plausibility (Kane, 2013). The more ambitious the claim, the more authority required of the warrants and the more evidence needed to support the warrants. Even if the claim is warranted by legitimate authorities, it may not hold true when the conditions of rebuttal apply. Rebuttals indicate circumstances under which the warranted conclusion is defeated or rebutted and may take the form of alternative explanations or counterarguments to the intended inference (Bachman, 2005). The strength with which the claim is made on the basis of data and in virtue of the warrants can constitute another component – Qualifiers. As such, qualifiers determine the degree of force conferred on the claim by the data (Toulmin, 2003). Since its inception, the argument structure has been influential in a variety of contexts including cognitive science, legal argumentation and educational measurement (Kunnan, 2010). Particularly, in the field of educational measurement, Kane (2013) drew on Toulmin’s argument structure to develop the argument-based approach to validation of the interpretation and use of test scores. The approach is discussed in greater detail in the following section.

3.2. The argument-based approach to test validation

The argument-based approach to test validation entails two stages closely related to each other: an interpretive argument which articulates the proposed interpretations and uses of test scores by specifying the claims, the network of inferences, and their associated assumptions; and a validity argument which comes later to evaluate the coherence, completeness, and plausibility of the interpretive arguments (Kane, 1992, 1994, 2012, 2013).

3.2.1. The interpretive argument

The interpretive argument entails three types of inferences leading from the observed performance to the conclusions and decisions based on the test scores, namely scoring inference, generalization inference, and extrapolation inference, each linked to one another in a systematic network (Kane, 1992; Kane, Crooks, & Cohen, 1999). Later, Bachman (2005); Bachman and Palmer (2010); Chapelle et al. (2008) adopted and expanded this framework in the context of language assessment to include two additional inferences: explanation inference and utilization inference. Figure 2.2 illustrates the network of inferences in the interpretive argument.

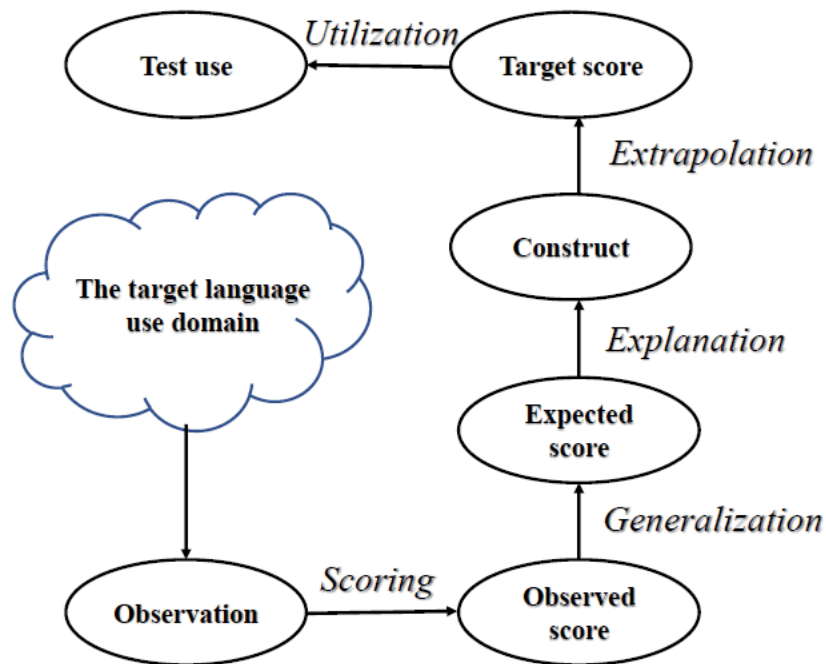


Figure 3. 2. The interpretive argument

Scoring inference

The interpretations and uses of test scores usually start with the elicitation of a sample of test-takers’ performance and the scoring of that performance on the basis of a predetermined scoring rubric or answer key. Therefore, the scoring inference provides a link between the observed performance as the data and the observed score as the claim. The scoring rule or rubric serves as the warrant for the inference. The warrant is backed by at least two assumptions: task administration conditions and the implementation of the scoring rubric have been consistently, appropriately and correctly delivered and managed; and in tests where human raters are involved, interrater and intrarater reliability has been assured.

Generalization inference

The interpretation and use of test scores based on a single delivery of a test form, in a particular context, and on a particular occasion usually attract little interest. Of more interest, however, is the possibility of generalizing test scores across multiple test administration conditions and with multiple test versions in a universe of generalization of which test performance can be taken as a sample and test scores can be considered as an estimated mean of the universe scores (Kane, 2012, 2013). The generalization inference illustrates the link between the observed scores as the data and the universe scores as the claim, the warrant of which is the consistency of test scores obtained over relevant parallel versions of tasks and test forms, and across raters (Chapelle et al., 2008). Backing for the generalization inference can be sought by ascertaining the representativeness of the sampling of tasks and conditions of observations as samples of the universe of generalization; and the elimination of any sampling errors that may bias one particular group of test-takers over another (Kane, 2013).

Theory-based/explanation inference

The interpretation of test score can also be attributed to a theoretical construct, which is defined by their roles in a theory (Cronbach & Meehl, 1955a). In language assessment, the role of construct is central to any discussion of language test validation (Chapelle, 2012b). In this connection, the interpretive argument offers “a more precise use for a construct definition than did previous formulations of validity argumentations” and “a more productive direction for discussion of language constructs” (Chapelle, 2012b, p.24). The theory-based inference, or what Chapelle (2008) termed *explanation inference*, links the observed score as the data to the claim about a theory-defined construct indicated by attributes about the kinds of observations relevant to the construct, the kinds of scoring rules used to evaluate those observations, and the conditions of observations that may affect the estimates of the construct (Kane, 2013). The degree of agreement between the indicators and the constructs determine the certainty with which claim about the theory-defined construct can be made. For example, if the theory-defined construct of academic language proficiency indicates that an academic reading test should measure three skills relevant to the domain of academic reading: reading for basic information, reading for general understanding and reading to learn, the underlying structure of the test should also reflect these indicators to the extent possible as measures of the theoretical construct. Backing for the

explanation inference includes evidence of the construct representation and lack of construct underrepresentation and construct irrelevant factors (Aryadoust, 2013).

Extrapolation

The interpretation and use of test scores for a particular observed performance should also extend to real-world performance as the observed performance neither comprehensively represents nor is taken as a random sample from the target domain. The extrapolation inference takes observed scores as data and the expected value over the target domain as the target score (Kane, 2012, 2013). The extrapolation inference rests on the warrant that the observed performance accurately predicts behavior in the real-world domain (Aryadoust, 2013), and thus should be supported by criterion-related evidence pertaining to the conventional concepts of predictive and concurrent validity. If the generalization inference can be considered as relevant within the test domain, the extrapolation inference take observed score beyond the test itself to the universe of real-world performance.

Utilization

Bachman (2005) and Bachman and Palmer (2010) introduced the utilization inference as an essential component of the interpretive argument because they believed that test uses and the consequences of test uses had been ignored in Kane's (1992, 1994) argument and that important decisions about the test-takers (e.g. whether they met the language requirements for university study) were made on the basis of the appropriate interpretation of test scores. The utilization inference, or decision inference (Kane, 2013), links target score as the data and the decision made on these scores as the claim. This inference is warranted by at least two assumptions: the cutoff scores set by the institutions for decision-making purposes actually reflect the proficiency levels of test-takers and full understanding and appropriate uses of these scores have been achieved by those who are involved in the decision-making process (Aryadoust, 2013; Chapelle et al., 2008).

The articulation of the interpretation and uses of test scores in an interpretive argument in terms of the inferences and their associated assumptions should be flexible and relevant to specific contexts and population to which it will be applied. Not all inferences and assumptions are relevant to every context and population, but rather some of them are plausible a priori while some are questionable. It is those critical or questionable inferences and assumptions that should receive most attention and need thorough empirical scrutiny. In any cases, however, the interpretation and use argument should be articulated clearly and in enough detail so as to allow for a transparent,

straightforward and well-informed examination of the evidence that support the claims being made, and to facilitate the evaluation of those claims in the validity argument stage.

3.2.2. The validity argument

“In the simplest term, a validity argument is an interpretive argument in which backing has been provided for the assumptions.”

(Chapelle, Enright, & Jamieson, 2010)

The validity argument involves the evaluation of the coherence and completeness of the interpretation and use argument (e.g. interpretive argument) and the plausibility of the inferences and assumptions (Kane, 2013). Kane (2012) suggested three criteria for the evaluation of the interpretive argument. First, the argument should be clear in the sense that the sequence of inferences, warrants, and assumptions are specified in a sufficiently meticulous manner so that the interpretation and use of test scores are well-informed in terms of what evidence is needed and how much is sufficient. Second, the argument should be coherent, illustrated by the persuasiveness inherent in the reasoning from the observed scores to the conclusion and decision made on the basis of the scores. Third, the inferences and assumptions should be tenable to relevant stakeholders. Although some assumptions are inherently reasonable and require little backing, most need support in terms of “careful documentation and analysis of procedures” and empirical evidence (Kane, 2012, p.13).

Different types of inferences require different kinds of analysis for their evaluation (Kane, 2013) and the field of language testing has documented various methodological techniques to examine the evidence required of each inference in the interpretive argument (Drackert, 2015).

Scoring inference

The scoring inference is based on two assumptions: appropriate use of scoring rules and consistent administration conditions. Appropriate use of scoring rules can be supported by experts’ judgment of the appropriateness of the scoring criteria or the scoring rubric development. In cases where human raters are involved, raters’ introspective and retrospective reports can be collected to examine if the scoring rubric has been consistently and effectively applied or even if the raters have been given proper training at all. Alternatively, measures of intra-rater and inter-rater reliability indices can be extracted based on statistical procedures to ensure that raters uniformly interpret the scoring rubric in its applications to particular performances, thus producing reliable scores. Consistent administration conditions can be backed by the identification and elimination

of extraneous factors to the test, such as environmental nuisances or cheating (Aryadoust, 2013; Bachman, 1990), or the standardization of the testing conditions.

Generalization inference

The generalization inference posits that the observed scores would be replicated over the universe of generalization which involves different facets of tasks, occasions, scoring rubric, particular items or raters. The generalization inference is considered valid to the extent that test items are representative of the universe of generalization and there are enough of them to statistically account for sampling errors (Kane, 2013), and that item difficulty varies with respect to test-takers' ability (Aryadoust, 2013). The generalization inference can be supported by reliability studies (Haertel, 2006), generalizability studies (Brennan, 2001), and more recently Rasch modelling (Bond & Fox, 2015), which yields indices of person reliability – the precision with which the test discriminates test-takers of differing proficiency levels as measured by their performance on the test – and item reliability – the extent to which items can be differentiated in terms of their difficulty (Aryadoust, 2013). In addition, Differential Item Functioning (DIF) can also be employed to examine if test scores can be generalized across different educational settings or with different L1 test-takers (Drackert, 2015).

Explanation inference

The explanation inference attributes the interpretation and use of test scores to a well-defined theoretical construct. Warrants for this inference can be supported by indications that the representativeness of the theory-defined construct has been reflected in the number, types and features of tasks in the test and the way these tasks are approached by test-takers; and the limitation of any factors other than the construct that may have an impact on the test performance and test scores. Several assumptions can be postulated to support this inference: a variety of linguistic skills, knowledge bases, and processes account for successful performance on the test; test item difficulty varies with respect to the intended attributes of test tasks; the test strongly correlates with other tests that measure constructs of the same types; and construct underrepresentation and construct irrelevant factors are well controlled. These assumptions can be endorsed by the analysis of test method effects, test dimensionality, cognitive processes, correlational relationship, item difficulty modelling, and differential item functioning (Aryadoust, 2013; Chapelle et al., 2008; Kane, 2004; Li, 2015a; Messick, 1989).

Extrapolation inference

The extrapolation inference connects test-takers' performance in the test domain to the real world activities, of which the test is a sample, by postulating that test scores accurately predict performance in the real world domain. This inference can be evaluated by the examination of the relationship between test scores and criterion-related scores that cover the target domain more thoroughly (Kane, 2013). This can be done through the correlational analysis of test scores and the test-takers' self-assessment of their ability, their grade-point average, or teachers' judgment of their ability (Aryadoust, 2013; Fan & Yan, 2017; Li, 2015a). The extrapolation inferences can also be evaluated by examining the extent to which the language elicited by test tasks in the test accounts for the language elicited in the target language use domains (Kane, 2013; LaFlair, 2017, Johnson & Riazi, 2017).

Utilization inference

The utilization inference concerns the precision and adequacy with which decisions about the test-takers are made based on their performance on the test. According to Kane (2013), the capstone of the interpretation and use argument for the score-based decisions is the decision rules which can be backed by the evaluation of how well these rules account for the target population and how effectively they limit the unintended or negative consequences. Several methods have been proposed to support this inference such as the use of Rasch measurement and expert judgment to evaluate if the test serves as a sufficient indicator for making judgment about test-takers (Aryadoust, 2013), the development and distribution of instructional and score interpretation materials, and washback studies (Chapelle et al., 2008).

3.2.3. Implementation of the argument-based approach

The argument-based approach to test validation has been widely adopted in various projects in language assessment. One of the earliest large-scale language assessment projects that employed this approach is the revision of the TOEFL test (the TOEFL iBT version) by Educational Testing Service. Drawing on their 7-year experience of involvement with the project (2000-2007), Chapelle et al. (2010) highlighted four key distinctions that gave advantages to the argument-based approach over the 1999 AERA/APA/NCME Standards for Educational and Psychological Testing (1999) (hereafter, the standards), which was prevalent at the time, and also explained why they adopted the argument-based approach for the TOEFL revision project rather than the standards.

First, the interpretive argument in the argument-based approach provided a more straightforward and explicit outline for framing the intended score interpretation than the standards

which is based on the definition and operationalization of a construct. Three reasons were given for the limitations of the construct definition approach: there had been no consensus on a definitive way to define the construct of language proficiency; limiting the construct of language proficiency to a particular aspect of knowledge (e.g. vocabulary, syntactic) ignored the complex processes and strategies involved in language processing, thus underrepresent the desired interpretation of the test; and confining the construct definition of language proficiency to a particular test obscured the nature of language performance which varies according to contexts of use. Instead of relying solely on the construct, the argument approach integrates the construct definition into a chain of inferences leading from the observed performance to generalization across the test domain, and ultimately to the target language use domain, thereby providing a more comprehensive account of the interpretation and use of the test scores.

Second, in terms of building a research design that can generate different types of evidence to support the interpretation and use of test scores, the argument-based approach has a sharper and stronger focus than the standards. The standards take the construct as the basis upon which different lines of evidence germane to the test content, response processes, internal structure, relations to other variables and consequences can be generated (Messick, 1989). In so doing, the validation process involves the consultation of a list of potential validity types without clear guidance as to which types should be discerned first, which later and how much of them is enough (Chapelle, 2012b). In contrast, the argument-based approach offers a coherent sequence of inferences with associated warrants and assumptions in the interpretive argument to facilitate a systematic examination of the validity evidence. As a result of this, the third distinction concerns the validity evidence derived from the validation research. While validity evidence generated as informed by the standards is just that – evidence – sporadically and separately situated within the construct validation paradigm, the evidence afforded by the argument-based approach can be structured into a coherent whole supporting the interpretive argument.

Finally, the specificity with which the interpretive argument is formulated lets it open the opportunities to challenge the validity argument in the forms of counterevidence or counterarguments as parts of the rebuttal conditions. The standards also allow for counterevidence, but this is confined to the construct per se in the forms of construct underrepresentation and construct irrelevant variance, thus representing a narrow focus of the whole testing cycle.

In the latest version of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, hereafter the 2014 Standards), the limitations associated the 1999 Standards as discussed above, were addressed by adhering to the tenets of the argument-based framework. The first and foremost change was the caveat against the use of the phrase “the validity of the test”. This is in line with the proposition of the argument-based framework that validity is not a property of the test itself, but rather the property of the proposed interpretation and use of the test scores, and that it is the proposed interpretation and use of the test scores that are evaluated, not the test. In addition, rather than taking construct definition as the starting point for the validation process, the 2014 Standards includes construct specification as a component in the proposed interpretation and use of the test scores, and “it is incumbent on test developers and users to specify the construct interpretation that will be made on the basis of the score or response pattern” (p.11). In terms of the validation procedure, the 2014 Standards suggests the accumulation of “types of validity evidence”, each tied to a specific proposition that underlies a proposed interpretation for a specified use. Since each test can be interpreted in multiple ways for multiple uses, multiple relevant sources of evidence would be needed in a programmatic research design. Such framing of “types of validity evidence” in the 2014 Standards offers a more systematic approach to the accumulation of evidence to support the intended interpretation of the test scores for a proposed use as opposed to the 1999 Standards.

The distinctions above underscore the advantages of the argument-based approach not only for the TOEFL revision project, but also for a variety of language test development and validation projects in different contexts (see for example, Aryadoust, 2013; Chung, 2014; Jia, 2013; Jun, 2014; Kadir, 2008; Llosa, 2008; Voss, 2012). However, apart from the TOEFL revision project which was a concerted effort of numerous researchers, the inclusion of every inference in the argument-based approach would be beyond the scope of a project carried out by an individual researcher. Instead, many of the studies that adopted the argument-based approach covered only one or several inferences in the interpretive arguments (Aryadoust, 2013; Jia, 2013; Li, 2015a), thus enabling a more in-depth inquiry into the claims that need more empirical justification. Therefore, the current study sets out to develop and evaluate the interpretive argument for the L-VSTEP reading test, focusing particularly on the explanation and extrapolation inferences which respectively attribute test scores to the theoretical construct of the L-VSTEP reading test and relate

students' observed performance on the test to their expected performance in the target language use domains.

3.3. Articulating an interpretive argument for the L-VSTEP reading test

3.3.1. Description of the L-VSTEP reading test

The L-VSTEP reading test measures test-takers' comprehension of four reading passages of 400 - 450 words each, through their responses to 40 multiple-choice questions (10 questions for each reading passage) in total. The difficulty level of the four reading passages and their associated comprehension questions is consistent with Levels 3 to 5 of the CEFR-VN. The reading passages represent topics in everyday life, natural, social, academic, and professional contexts, which require no specialized knowledge or experience for comprehension. As illustrated in the test specifications (MOET, 2015b), each L-VSTEP reading test assesses a range of different subskills, including understanding explicit information (eg., meaning directly presented in the texts, meaning directly presented in the texts but phrased differently), understanding cohesive devices (eg., understanding text coherence based on connective devices or references), inferring meaning of unfamiliar words (e.g., inferring meaning of unknown words from contexts), inferring situational meaning (e.g., making inferences based on contextual clues), integrating textual information (e.g., synthesizing information across the text), understanding pragmatic meaning (e.g., understanding author's purposes, attitude, and stances), summarizing textual information (e.g., understanding main ideas of a text), and recognizing text structure (e.g., appreciating the organization of a text) (Nguyen, 2018).

The test is delivered on a paper-and-pencil format and is manually scored by examiners. Test takers' performance is scored on a 10-point scale which is then divided into four score ranges. Those who have 32 – 40 correct items fall into the 8.5 – 10 range (corresponding with Level 5 of the CEFR-VN), 19 – 31 in the 6 – 8 range (Level 4), 11 – 18 in the 4 – 5.5 range (Level 3), and 0 – 10 in the 0 – 3 range (under-achieving). The score range for the reading component was decided on the basis of the results of pilot tests and are illustrated in Table 3.1.

Table 3. 1. Score range for the VSTEP reading test

CEFR-VN levels	Number of correct items	Score range
Under-achieving	0 – 10	0 – 3
3	11 – 18	4 – 5.5

4	19 – 31	6 – 8
5	32 – 40	8.5 – 10

The VSTEP manual (ULIS, 2015) offers written description of test-takers' reading proficiency based on their test results and in accordance with Levels 3 to 5 of the CEFR-VN. Details are presented in Table 3.2.

Table 3. 2. Description of reading proficiency based on L-VSTEP reading test scores

Level 5 (8.5 – 10)	
Test-takers are able to understand details of different types of reading which vary in length and content in everyday life, social, professional and academic situations; are able to recognize author's attitude and opinions, understand implicit and explicit information.	Specifically, test-takers at this level are able to recognize purposes and arguments of the author, are sensitive toward cultural elements of English embedded in the text, able to understand a variety of idiomatic expressions and to recognize structures of information and logical development of complex text through the analysis of text organization
Level 4 (6 – 8)	
Test-takers are able to comprehend different kinds of texts such as news reports, newspaper articles, work reports, etc. in a variety of professional contexts.	Specifically, test-takers at this level have a wide vocabulary, but may still have difficulty with less familiar idiomatic expressions. New words may not interfere with their comprehension of texts thanks to their ability to guess word meaning from context or skip unimportant words. They are able to skim the texts for general information, understand rephrased information, or implicit details such as author's attitude, opinions, style, though with some difficulty.
Level 3 (4 – 5.5)	
Test-takers are able to understand explicit information in texts about topics of interest.	Specifically, test-takers at this level are able to understand/recognize important/explicit information in texts about familiar topics,

able to skim through long texts, locate important information via connecting devices or referenced words and gather information from different parts of a text or from different texts for a specific task. They are also able to recognize argument structures of a text, though with difficulty.

The above description of the L-VSTEP reading test highlights several assumptions that inform the articulation of the proposed interpretation and use of the test scores. First, since the test is intended to assess a variety of reading subskills, these intended subskills should be identifiable from students' performance on the test via either the underlying structure of their test scores or the cognitive processes through which they answer the reading items. Second, because the difficulty of the reading texts and reading comprehension questions of the test is anchored at levels 3 to 5 of the CEFR-VN, it is important to empirically ascertain that the complexity and difficulty level of the linguistic features of the reading texts and reading comprehension items is consistent with the item difficulty of the test as informed by those students whose performance on the L-VSTEP reading test falls within levels 3 to 5 of the CEFR-VN. Finally, since students' scores on the test reflect their reading proficiency at level 3 to 5 of the CEFR-VN as described in the test specifications, there should be empirical evidence on the alignment between students' performance on the test and their performance in the target language use domain that is consistent with reading proficiency described in the CEFR-VN levels 3-5. The following sections, therefore, offer a detailed discussion of the explanation inference and extrapolation inference of the argument-based framework which draw on these assumptions of the intended construct of the L-VSTEP reading test.

3.3.2. The explanation inference

Since the explanation inference relates test scores to a theoretical construct which is fundamentally analogous to the conventional concept of construct validity, it is essential that the formulation of the assumptions that underpin this inference be sufficiently informed by influential perspectives about construct validity. In this connection, two prominent perspectives about

construct validity should be taken into account, namely the nomological network (Cronbach & Meehl, 1955a) and the nomothetic span (Embretson, 1983).

The nomological network

According to Cronbach and Meehl (1955a), construct validation is conducted when the investigator believes that “his instrument reflects a particular construct, to which are attached certain meanings” (p. 290). Therefore, construct validation is basically a theory testing process during which theories or hypotheses about a latent attribute or trait are postulated, and then confirmed or disconfirmed by observable performance. This process must be situated within a nomological network which is characterized as a network of laws specifying the relations among (a) observable quantities or properties to each other; (b) theoretical constructs to observables; or (c) different theoretical constructs to one another (p. 290). The relation between constructs and observations in the network is critical because if observations do not fit in the network, construct validation cannot be claimed and modifications of the network or alternative constructs must be made. Several methods for construct validation have been proposed: for example, group difference can be tested if an English test is constructed to differentiate learners of lower and higher English proficiency levels. Alternatively, the internal structure of a test can be examined if it is hypothesized that the trait being measured requires significant intercorrelation among certain items in the test, and non-significant or negative correlation among other items irrelevant to the postulated construct. Studies of processes should also be conducted to identify what precisely accounts for variability in test scores. For instance, reading comprehension test scores would be interpreted differently from the postulated constructs if it is found that failure to deliver an item correctly is attributed to misunderstanding of the question stem rather than to the lack of understanding of the passage.

The nomothetic span

Unlike Cronbach & Meehl’s (1955a) description of the nomological network which involves the identification of the theoretical mechanisms such as information processes, strategies and knowledge bases that underlie task performance within a single test, Embretson (1983), under what she called a paradigm shift from functionalism to structuralism, proposed an additional approach to construct validation – the nomothetic span. The nomothetic span denotes another network of relationships between the scores of a test with external variables such as other test measures that share similar constructs or measures of the same constructs under different

circumstances. A notable distinction between the nomological network and the nomothetic span, according to Embretson, lies in the approaches to research that they induce. While the nomological network involves construct representation research that deals primarily with task variability rather than person variability, nomothetic span is primarily concerned with the importance of tests as measures of individual differences (Embretson, 1983, p.180). That said, the introduction of the nomothetic span is not at odds with the nomological network, but rather expands on the latter by incorporating both the correlation between the test and the target latent traits, and between the test and other tests within a unified paradigm (Aryadoust, 2013). Construct validation studies within the nomothetic span should take into account four issues: test dimensionality should be meaningful as indicators of component abilities (e.g. person measurements); component abilities should exhibit external validity as measured by its degree of correlation with other criterion measures; component abilities should show differential validity as measured in different learning environments or different content areas; and the component abilities should indicate across-task generalizability.

Three planes of explanation

Drawing on the above conceptualizations of construct representation and nomothetic span, and the concrete approach to the analysis of task characteristics, Chapelle et al. (2008) proposed a model for articulating and evaluating the explanation inference in the argument-based approach which consists of three planes, or strata, connected to each other in a hierarchical order of abstractness (Figure 2.3).

At the highest level of abstractness lies the plane of language proficiency construct. It is abstract in the sense that the construct defined in this plane refers to general language proficiency rather than focuses on any particular knowledge aspects or processes pertaining to a specific sample of performance. This plane implies several approaches to research in search of evidence as backing for the explanation inference, such as the examination of how the amount and quality of English learning experience accounts for variability in test performance; the investigation of the correlation between test-takers' performance on the test and performance on other test measures concerning the nomothetic span; and the identification of the internal structure of the scores as highly intercorrelated components explaining a priori theoretical expectations.

The middle plane represents constructs in a relatively more concrete sense, which specifies the nomological network of psycholinguistic knowledge, processes, and strategies involved in test

performance. Research informed by this construct representation approach entails the examination of the specific cognitive processes, knowledge bases and strategies used by test-takers while taking the test and the extent to which these components reflect theoretical expectations.

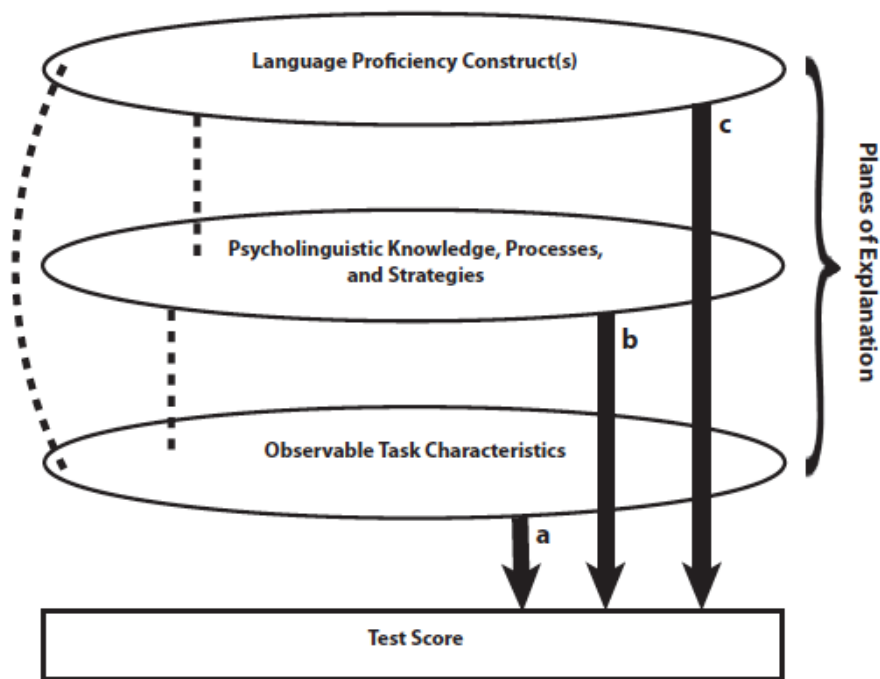


Figure 3.3. The three-plane model of explanation inference (Chapelle et al., 2008, p.336)

The most concrete plane shapes construct definition in terms of the concrete surface features of tasks that explain performance consistency such as the number and types of words in a text, the number of clauses in a sentence, the number of distractors in a multiple choice item, or the percentage of overlap between question stems and the texts.

In Figure 3.3, the three arrows pointing from (a) observable task characteristics, (b) psycholinguistic knowledge, processes and strategies, and (c) language proficiency construct toward the test score represent the extent to which the three planes provide explanations for test scores as informed by the evidence yielded in the validation research. The dotted lines illustrate an ideal connection among the three planes that provides the links for an integrated construct theory (Chapelle et al., 2008). This model offers an explicit and straightforward basis to articulate an explanatory interpretation and use argument for the VSTEP reading test.

The explanation inference for the L-VSTEP reading test

Since the explanation inference relates the L-VSTEP reading test scores to a theoretical construct of general English reading proficiency, the claims and assumptions which are associated

with the explanation inference within an argument-based framework, and which are informed by the model of explanation (Chapelle et al., 2008) are proposed for the interpretive argument in the current project. Table 3.1 summarizes the explanation inference for the L-VSTEP reading test in terms of the assumptions and corresponding research questions, the backings and rebuttals.

The three assumptions which are proposed to authorize the explanation claim of the L-VSTEP reading test, and which correspond with the three levels of abstractness of the explanation planes include: (a) the observable linguistic and discourse characteristics of the texts, items, and item-text interaction explain item difficulty (RQ3); (b) the reading processes and strategies employed by test-takers conform to theoretical expectations (RQ1); and (c) the internal structure of the L-VSTEP reading test reflects highly intercorrelated components explaining theoretical expectations (RQ2). In order to yield sufficient evidence to back these three assumptions, three different lines of research inquiry are also suggested: (a) the analysis of linguistic and discourse features of text and items using text analysis software and the statistical modelling of their contribution to item difficulty; (b) the use of stimulated verbal recall to gain insights into test-takers' thought processes and strategies while reading; and (c) the use of factor analysis to identify the theory-informed factor structure of the test. The strength with which the explanation inference claim is made increases if it survives the rebuttals. Potential rebuttals to the explanation inference of the L-VSTEP reading test may include: (a) item features (e.g. the question stem contains high percentage of low-frequency vocabulary) overshadow text features in explaining item difficulty; (b) test-wiseness strategies or guessing are used to answer the questions; and (c) no discernable underlying patterns of test scores are recognized.

Table 3. 3. Summary of the explanation inference for the L-VSTEP reading test

Research question	Warrant	Assumption	Backing evidence	Potential rebuttal
1. What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes	1. Students' scores on the L-VSTEP reading test can be attributed to	1. Reading processes and strategies engaged by test-takers vary according to	1. Stimulated verbal recall of test-takers' reading processes and strategies while doing the L-VSTEP reading test	1. Test-takers might employ test-wise strategies and wild guessing to answer the test questions.

correspond with the reading processes actually engaged in by test-takers while doing the test?	the construct of English reading proficiency	theoretical expectations	and expert judgment on the skills and strategies elicited by the test items	
2. To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with different reading proficiency levels and academic disciplines?		2.a. The internal structure of test scores reflects highly intercorrelated components explaining theoretical expectations (b) Internal structure of test scores remains invariant across different groups of test-takers	2.a. Factor analysis of the underlying structure of the L-VSTEP reading test (b) Test of measurement invariance of the underlying structure of the L-VSTEP reading test	2.a Large amount of unexplained residues (b) No discernable underlying patterns are recognized (c) Internal structure of test scores is non-invariant.
3. What are the linguistic and discourse characteristics of the texts, items and item-by-text variables of the L-VSTEP reading test? To what extent do		3. Observable task characteristics underlie task performance consistency	3. Analyses of texts, items and item-by-text linguistic and discourse characteristics Multiple regression analyses of linguistic and	3. Features of items and item-by-text override features of texts in explaining item difficulty

they contribute to item difficulty?	discourse features as predictors of item difficulty
--	---

While the explanation inference provides a framework to relate students’ performance on the test to a theoretical construct of language proficiency, it does not account for the extent to which students’ performance on the test can be generalized beyond the test domains and into the real-world target domains. The extrapolation inference is, therefore, proposed to elucidate this real-world generalization.

3.3.3. The extrapolation inference

In their search for a unified conceptualization of language proficiency to serve as a basis for test design and score interpretation, Chapelle et al. (2008) proposed a dual ground basis which incorporated the competency-centered perspective and the task-centered perspective. The former relates score interpretation to a theoretical construct of language ability, which constitutes the explanation inference while the latter sets the context of language use as a basis for score interpretation. The relevance and importance of context of language use, as they maintained, was brought to bear on the one-size-fits-all belief associated with the competency-centered perspective that failed to account for variation in language performance across different contexts. Therefore, from a task-centered perspective, the type of tasks considered to be important in the real-world context must be identified and approximated as much as possible in the test domain. Understanding the contextual characteristics of the target language use domains and designing tasks that approximate those characteristics to elicit comparable samples of language performance in the test engendered evidence that supports the domain description inference during test development in the argument-based framework. However, this process can also be conducted as a post-hoc evaluation of the comparability between the test tasks and the target domains to support the extrapolation inference (LaFlair & Staples, 2017).

The extrapolation inference in the current project links observed test scores with the expected scores in the target domains by postulating that students’ observed performance on the L-VSTEP reading test predicts their expected performance in the target language use domains. Since the primary purpose of the L-VSTEP test as investigated in the current project is to serve as

a screening tool for tertiary students who seek to achieve a minimum standard of English proficiency prior to their graduation, the target domains in which language use occurs are the academic programs that they are pursuing at the relevant institution. The warrant for this inference can be premised on two mutually related assumptions. First, language performance in the test and language use in the target domains are subject to common contextual features (LaFlair & Staples, 2017). More specifically, the alignment between the reading tasks and skills required in the academic programs and those sampled in the L-VSTEP reading test is established. To the extent that the range of reading skills and tasks in the test covers most of the reading tasks and skills in the academic domains the extrapolation inference can be considered plausible. By contrast, if the reading tasks and skills included in the test substantially depart from the academic domains the extrapolation inference is weakened. The second assumption is that students' observed performance on the test predicts their expected performance in the target language use domains. In other words, there is a predictive relationship between students' scores on the L-VSTEP reading test and the expected scores that represent a more comprehensive assessment of their reading performance in the academic programs.

Backings for the extrapolation inference, therefore, can be sought via both analytic and empirical evidence (Kane, 2013). Analytic evidence involves the exploration of the perceptions and experience of the key stakeholders, namely lecturers and graduate students, who have experience with both the academic domains and the test domains in terms of the commonalities between the reading tasks and skills sampled in the test and those considered important and required in the academic domains. Empirical evidence can be generated by the study that statistically models the predictive relationship between students' scores on the test and their expected scores in the target domains that represent their reading ability more thoroughly. The latter can be yielded by students' self-reported English reading proficiency via a self-assessment questionnaire. Table 3.2 summarizes the extrapolation inference with its warrant, assumptions, backings and potential rebuttals as well as the relevant research questions.

Table 3. 4. Summary of the extrapolation inference for the L-VSTEP reading test

Research question	Warrant	Assumption	Backing evidence	Potential rebuttal
4. To what extent do students'	The observed performance	1. Students' test scores	1. Statistical modelling of the	1. No relationship

<p>scores on the L-VSTEP reading test predict their performance in the relevant academic programs?</p>	<p>of students on the L-VSTEP reading test relates to their English reading performance in the academic programs at the relevant institution.</p>	<p>significantly predict their performance in the academic programs as assessed by their self-reported English reading proficiency.</p>	<p>predictive relationship between students' test scores and their self-reported reading proficiency</p>	<p>between students' test scores and their self-reported English reading proficiency is found.</p>
<p>5. To what extent are reading tasks and skills assessed in the L-VSTEP reading test aligned with reading tasks and skills required in the relevant academic programs?</p>	<p>academic programs at the relevant institution.</p>	<p>2. The reading tasks and skills as assessed in the L-VSTEP reading test are compatible with those required in the relevant academic programs.</p>	<p>2. Examination of the lecturers' and graduate students' perceptions about the commonalities between reading tasks and skills sampled in the test and those required in the academic programs.</p>	<p>2. There is an underrepresentation of test tasks and reading skills in the test as compared with those in the academic domains.</p>

CHAPTER IV: METHODOLOGY

This chapter delineates the research methods employed to answer the study's research questions. As discussed in chapter II, a mixed-method approach can provide an essential springboard for the current study to examine L2 reading both as a product and as a process. The present chapter first provides a brief introduction to mixed method design with its tenets, principles, and relevance to the study. This is followed by a discussion of the sample and sample size requirements. A detailed discussion of the data collection procedure and data analysis methods for each of the research questions constitutes the major part of the chapter.

4.1. Mixed methods research

Mixed methods research is defined as an inquiry approach that collects, analyzes, and combines both qualitative and quantitative data, concepts, techniques and language within a single study or inquiry program (Creswell, 2012; Creswell & Clark, 2018; Jang, Wagner, & Park, 2014; Johnson & Onwuegbuzie, 2004). It is promoted as a response to the schism between quantitative and qualitative research paradigms which Johnson and Onwuegbuzie (2004) referred to as the “paradigm wars” and which Teddlie and Tashakkori (2003, p.14) termed “the incompatibility thesis”.

Traditionally, advocates of the quantitative paradigm hold a strong belief adhering to the postpositivist philosophy with the ontological assumption that realities or truths exist independently of the human mind. Therefore, researchers engaging in a research program must have their emotion, biases, and subjective judgment removed, and replaced by a hypothesis testing process where validity, reliability, objective and empirical justifications play central roles. This research paradigm is characterized as consisting of deductive reasoning, hypothesis testing, prediction, confirmation, and generalization (Johnson & Onwuegbuzie, 2004; Moeller, Creswell, & Saville, 2016). In contrast, qualitative purists contend that constructivism rather than postpositivism should serve as the philosophical basis for the research inquiry practice. This philosophical basis posits that reality is subjective, and therefore, subject to multiple interpretations depending on multiple perspectives each individual brings to the world. As such, the purpose of this inquiry approach is to produce inductive reasoning, discovery, exploration, theory generation, and subjective judgment, which yields a richer understanding of the socially-

constructed and value-bound context (Johnson & Onwuegbuzie, 2004; Moeller et al., 2016). The polarization of these two research paradigms engenders a seemingly incompatible position where advocates of each paradigm criticize those of the other in defence of their philosophical stance (Riazi & Candlin, 2014).

Mixed method research represents an alternative paradigm to the traditionally-held paradigmatic polarization (Jang et al., 2014) by incorporating a pluralist worldview whereby flexibility in philosophical assumptions allows for the manipulation of multiple methods to answer the research questions at hand (Moeller et al., 2016). By carrying flexible philosophical assumptions (pragmatism, dialectical, and transformative), mixed methods research can serve different research purposes (triangulation, complementarity, development, and initiation) in a flexible manner, which in turn, is conducive to flexible research designs (convergent, exploratory, or explanatory) (Riazi & Candlin, 2014). For example, a pragmatism assumption places emphasis on the practical issue of successfully addressing a research problem given a research context, resources and the repertoire of research methods available at the researchers' disposal. They may combine different methods to confirm or cross-validate findings within a triangulation inquiry. And in so doing, they may employ a convergent design where data are generated independently but are then merged and compared concomitantly (Jang et al., 2014). The methodological approach of the current study is framed within the pragmatism paradigm.

The field of language testing and assessment has witnessed a paradigm shift from a strong focus on quantitative approach to a more balanced and flexible mixed methods approach in response to the ever changing and expanding definition and conceptualization of language competence and test validity (Jang et al., 2014). Specifically, this shift in paradigmatic perspectives is driven by two main reasons. First, the chronological evolvement of the concept of validity (as summarized in chapter I) and the shift of focus from the validation of test scores to the validation of the proposed interpretation and use of test scores in language test validation (as illustrated in chapter III) have inevitably given rise to the use of multimethod in lieu of monomethod approaches to address the multifaceted nature of language testing and assessment including, inter alia, the issues of social practices, curricular reforms, policy making, and accountability. Second, language ability as a key construct in language testing and assessment lends itself to various interpretations, definitions, conceptualizations and operationalizations, all of which require researchers to situate themselves within a dynamic and interacting context of “linguistic, psychosocial, political, and

cultural dimensions of language competence” (Jang et al., 2014, p.124), and neither of which, therefore, can be investigated from a single worldview. The paradigmatic shift in language testing and assessment has materialized itself at all levels of the inquiry system, from the social impacts of language assessment associated with the concept of washback to the validity of the interpretation and use of test scores.

Validation of the interpretation and use of the L-VSTEP reading test – the central focus of the current study - is contextualized within this dynamic paradigmatic movement. As reviewed earlier, L2 reading is a multicomponent, multiprocess construct. It is conceptualized from both product and process perspectives. While the product perspective places emphasis on the score such as score patterns, item difficulty and test-takers’ relative performance on the test, the process perspective attends to the actual reading processes that are conducive to the test scores. Neither of them alone can account for the whole enterprise of reading, and therefore should be integrated in a mixed method paradigm where understanding of one facet supplements that of the other. In addition, the argument-based approach, particularly the model of explanation inference and the extrapolation inference (Chapelle et al., 2008) adopted in the current study specifies that the construct of the L-VSTEP reading test be subject to empirical justifications pertaining to the reading processes and strategies, the surface features of the test tasks, the general concept of reading proficiency, and the contextual features that affect both the test and the target language domains, all of which require multiple approaches to data collection and analysis.

A convergent parallel mixed methods design (Creswell, 2012) was employed to address both the product and process conceptualizations of the construct of the L-VSTEP reading test and the comparability between the test and the target language use domains. The quantitative approach helps identify the underlying structure of the test, the test item difficulty, the multiple text features that explain item difficulty, and the predictive model of test scores and performance in target language use domains while the qualitative approach unpacks the underlying reading processes of the readers while they take the test as well as relevant stakeholders’ perceptions about the comparability between the test and the target language use domains. Results are then compared to examine the extent to which the qualitative data converge or diverge from the quantitative data as well as the sources of the convergence or divergence. Such a mixed methods design enables the investigation of the five research questions:

1. What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes correspond with the reading processes actually engaged in by test-takers while doing the test?

2. To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with different reading proficiency levels and different academic disciplines?

3. What are the linguistic and discourse characteristics of the texts, items and item-by-text variables of the VSTEP reading test? To what extent do they contribute to item difficulty?

4. To what extent do students' scores on the L-VSTEP reading test predict their reading performance in the relevant academic programs?

5. To what extent are reading tasks and skills assessed in the L-VSTEP reading test aligned with reading tasks and skills required in the relevant academic programs?

Table 4.1 illustrates the mixed methods design used in the study.

Table 4. 1. The convergent parallel mixed methods design

RQs	Data collection	Data analysis	Methods	Validity argument
1. What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes correspond with the reading processes actually engaged in by test-takers while doing the test?	1. Expert judgment 2. Stimulated verbal recall	Thematic analysis	Qualitative	
2. To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with different reading proficiency levels and different academic disciplines?	Students' test scores	1. Confirmatory factor analysis 2. Rasch analysis 3. Measurement invariance	Quantitative	Explanation inference
3. What are the linguistic and discourse characteristics of the texts, items and item-by-text variables of the VSTEP reading	1. Linguistic and discorsal features of the reading passages	1. Automatic textual analysis	Quantitative and qualitative	

test? To what extent do they contribute to item difficulty?	2. Students' test scores	2. Correlational analyses 3. Multiple regression analysis 4. Expert judgment 5. Rasch analysis		
4. To what extent do students' scores on the L-VSTEP reading test predict their reading performance in the relevant academic programs?	1. Students' test scores 2. Students' self-reported English reading ability.	1. Confirmatory factor analysis 2. Structural equation modelling	Quantitative	
5. To what extent are reading tasks and skills assessed in the L-VSTEP reading test aligned with reading tasks and skills required in the relevant academic programs?	1. Semi-structured interviews with lecturers and graduate students	Thematic analysis	Qualitative	Extrapolation inference

4.2. Participants

4.2.1. Sampling methods and participant recruitment

Sampling methods

Since the VSTEP was made official across the country in 2015, it has been used as proof of English proficiency for graduates by many tertiary institutions, including UA – the pseudonym of the university that the researcher has connection with, and hence has the potential to recruit enough participants for the study. Therefore, convenience sampling which is used when participants are willing and available for data elicitation (Creswell, 2012) was employed for the recruitment of participants. Table 4.2 illustrates the sampling of participants in the current project. Detailed criteria for participant recruitment and recruitment procedure are discussed in the following section.

Participant recruitment

Students

The primary purpose of the L-VSTEP reading test as conceptualized in the current study is to serve as proof of English reading proficiency for university graduation. Therefore, the participants recruited in the study are expected to represent, as closely as possible, those who are planning to sit the L-VSTEP test for graduation purpose. Two additional criteria are set for the recruitment. First, participants are expected to represent a wide variety of English language proficiency because the VSTEP test is designed to target learners at different proficiency levels from level 3 to level 5 of the CEFR-VN (see chapter I). One indicator for English language proficiency can be participants' English reading scores in their latest end-of-semester exam. Second, participants should represent a wide range of academic disciplines since the L-VSTEP is used for students across different academic disciplines rather than for English major students only.

Consent forms, with details of the objectives of the research project, participants' involvement and the potential benefits, were sent to students who meet the recruitment criteria at UA to which the researcher is granted access. Those students who agreed to participate were sent an official invitation with information about the next steps. In total, 544 students were recruited for the project. The recruited 544 participants were administered version A of the test to generate data for answering RQ2. Immediately following this administration, 9 out of these 544 participants representing three levels of English proficiency in reading (based on their scores on the test version A) were invited to participate in the stimulated verbal recall to provide data for RQ1. To generate data for RQ3, the 544 students were divided into four groups, each consisting of 136 students. Division was based on their scores on version A of the test so that each group consisted of an equal number of students at three different levels of reading achievement (levels 3, 4 and 5 of the CEFR-VN as detailed in chapter I). Three of the four groups were randomly assigned to take either version B, C or D of the test, four weeks after the administration of the test version A. The same 544 students who took the test version A also answered a self-assessment questionnaire of English reading proficiency to generate data for RQ4. The self-assessment questionnaire was completed before students took version A of the test to reduce the impact of test taking on their self-perception about English reading ability.

Experts

The literature on reading assessment has documented the use of expert judgment as a potential method for examining different aspects of test validity, such as to predict item difficulty and item discrimination of a test (Bejar, 1983; Choi & Moon, 2020; Fulcher, 1997), to identify the

reading skills targeted by reading test items (Alderson & Lukmani, 1989b; Anderson, 1990; Dawadi & Shrestha, 2018), and to judge the comparability between reading tests in terms of their construct, content, and task characteristics. Major concerns regarding the use of this method, as cautioned by previous scholars, involve the consistency and accuracy with which expert judgments can be made (Alderson & Lumley, 1995; Dawadi & Shrestha, 2018). Therefore, necessary conditions should be provided, and quality control protocols should be exercised to facilitate experts' judgment of the reading processes required by the test items in this study. Following Alderson and Lumley (1995), the following protocols were strictly adhered to during the data collection and analysis.

- 1) Criteria for recruiting experts should be clearly delineated and followed to make sure that experts are comparable in terms of theoretical and practical backgrounds and knowledge.
- 2) Reading skills as specified in the test development guidelines should be described clearly and succinctly in the expert judgment form to facilitate experts' understanding and judgment.
- 3) Training sessions should be conducted carefully and thoroughly to familiarize experts with the whole process of judgment and to rule out any misconceptions and disagreement about the interpretation of the skill descriptions that might later threaten the consistency and accuracy of the main judgment session.

As argued by Brown and Hudson (2002), the accuracy of the judgment can be affected by the professional views held by the experts toward the task at hand. In the current study, for example, experts might produce conflicting judgments if they have different beliefs concerning the divisibility of reading subskills, or even doubt whether subskills do exist at all. Therefore, to preclude any disparity between the experts in terms of their theoretical and practical beliefs, backgrounds and knowledge that might undermine the accuracy of the judgment, clear and well-defined criteria for participant recruitment should be formalized and followed. The following criteria were, therefore, adopted for the recruitment of expert participants. First, they should have at least a master's degree in TESOL or related areas where, in the study program, they have finished a course in language testing and assessment. Second, they should be actively involved in the development and evaluation of second language reading tests in their own teaching practices. Third, they should be familiar with the test development guidelines for the L-VSTEP reading test.

Finally, they should have experience teaching L2 reading courses to undergraduate students at their own institution.

Two language testing experts were invited to participate in the project. Experts' involvement in the project entailed three stages. First, they were required to code the item-text variable of plausible distractor in all four versions of the test (discussed in section 4.3.3) to generate data for RQ3. Second, they provided judgment of the reading processes intended to be measured by items in the test version A to address RQ1 (discussed in section 4.3.1). Finally, the same judgment data from the previous stage were used to build the Confirmatory Factor Analysis models to answer RQ2 (discussed in section 4.3.2).

Lecturers and graduate students

Three university lecturers and three newly graduated students at UA were invited to participate in individual semi-structured interviews. Each of them shared their perceptions about the alignment between the academic programs at UA and the test in terms of the reading tasks and skills required. These lecturers and students were selected on the premise that they had sufficient experience with and knowledge of the kind of reading tasks and activities normally encountered in the academic domains, the level of English reading proficiency required to function properly in those domains, and the L-VSTEP reading test format.

Table 4. 2. Sampling of participants in the project

RQs	Test versions	No. of students	No. of experts	Sampling methods	Activities
RQ1	A	9 (out of 544)	2	Convenience	Nine students participated in a stimulated recall section after completing version A of the test to report on their reading processes while doing the test Two experts provided judgments of the reading processes deemed to be elicited by the items in version A of the test
RQ2	A	544	2	Convenience	544 students took version A of the test to provide data for

					confirmatory factor analysis of the test's underlying structure Two experts assigned items from the test version A to their underlying factors.
RQ3	A, B, C, D	544	2	Convenience	544 students were divided into four groups (136 students each). Three random groups out of four took either version B, C or D of the test to generate item difficulty measures. Two experts coded the item-text variable of plausible distractor
RQ4	A + question naire	544	0	Convenience	544 students who took version A of the test answered a self-assessment questionnaire of reading proficiency.
RQ5	0	3	3	Convenience	Three lecturers and three graduate students took part in semi-structured interviews.
Total	4	544	5		

4.2.2. Sample size requirements

Quantitative and qualitative research methods have different requirements regarding sample size. Since quantitative methods aim to test hypotheses, confirm or disconfirm theories, and generalize results, sufficiently large sample size is essential to control sampling errors, deal with missing data, eliminate extreme cases, and satisfy statistical assumptions. On the other hand, qualitative methods focus on explanation, exploration, subjective judgment, and personal viewpoints to yield a rich and thorough understanding of the research problems. Therefore, it is a common practice to select a small number of participants and delve into each and every aspect of

research interest that they have to offer, thereby extracting the most relevant and elaborate set of data for interpretation. The required sample size is reported below in relation to each RQ.

In order to address RQ3, two major statistical procedures were employed: multiple regression analysis and Rasch analysis. The simplest rule of thumb for estimating the sample size for multiple regression analysis is to have 10 to 15 cases per predictor variable (Field, 2009). For example, if there are 12 predictor variables in the regression analysis to be performed in the study the estimated sample size is in the range of 120 – 180. Alternatively, Green (1991) proposed two formulae to calculate the sample size required for multiple regression analysis. If the priority is to test the overall model fit, the formula is $50 + 8k$ where k is the number of predictor variables. If the relative importance of each individual predictor is prioritized, the formula is $104 + k$. This study adopts the latter formula because it aims to identify which of the predictor variables account for most significant variance in the item difficulty.

Regarding regression analysis, the predictor variables of the study are the text, item, and item-text features while the criterion variable is the item difficulty of the test. Therefore, the sample size does not apply to the “test-takers” that participate in the study as is normal the case, but rather the number of items of the L-VSTEP reading tests. As the L-VSTEP consists of 40 items, a decision was made to include four practice versions of the L-VSTEP, namely versions A, B, C, and D. In this way, a sample of 160 items from the four versions of the L-VSTEP reading test met the minimal sample size requirement discussed earlier. The four practice versions of the L-VSTEP reading test were obtained from the official training package of the test with consent from the test developers at UA. The four test versions are considered equivalent on the basis that they are developed out of the same test development guidelines, and by the same team of test designers. This is essential for the linking of dataset that allows for comparisons across test forms (Barkaoui, 2015).

The determination of the sample size for the dichotomous Rasch model depends on numerous factors such as “number of items, location of items along a trait, overlap of items along a trait, distribution of respondents along a trait, number of respondents, targeting of items to persons along a trait, and the goal of an instrument” (Boone, Staver, & Yale, 2014). However, there are several rules of thumb suggested by previous researchers. For example, Wright and Tennant (1996) argued that “with a reasonable targeted sample of 50 persons, there is 99% confidence that the estimated item difficulty is within +/- 1 logit of its stable value” (p.468).

Bamber and van Santen (1985) suggested that since we are more concerned about item fit than person fit, there must be more persons than items. So, for a 100 – item dichotomous test, we would need a sample of 100+ persons.”

RQ1 and RQ5 are informed by a qualitative approach to data collection and analysis. Therefore, a small sample of participants (nine and six for RQ1 and RQ5 respectively) who represent the targeted population were recruited so that indepth insights into their reading processes while taking the test and their perceptions about the test and the relevant academic domains can be exploited.

The statistical methods used to address RQ2 and RQ4 are Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM) which are based on large sample theory (Lehmann, 1999). The minimum required sample size for CFA/SEM is affected by various factors such as the multivariate normality of the data, the estimation method, model complexity, and the amount of missing data (Hair, Black, Babin, & Anderson, 2014). Following the suggestions of Cohen (1992), and Bentler and Chou (1987), this study employed a sample of 544 participants which suffice to render the statistical modelling methods of CFA/SEM and measurement invariance viable.

Detailed description of the whole data collection and data analysis procedure to answer each of the research questions is presented in the following sections.

4.3. Data collection and data analysis procedure

4.3.1. Research question 1

What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes correspond with the reading processes that are actually engaged in by test-takers while doing the test?

This research question addresses the middle plane in Chapelle et al. (2008) explanation model that concerns the examination of the test response processes and strategies and the extent to which these processes and strategies vary with respect to a theory-defined construct. More specifically, it seeks to establish the correspondence between the reading processes assumed to be instigated by the L-VSTEP reading test items and the actual processes employed by the test-takers while taking the test. In order to inform the research design process, two assumptions are made with regard to this correspondence. First, the cognitive processes employed by test-takers while answering a particular L-VSTEP reading test item should approximate those processes deemed by

experts to be instigated by that particular item. Second, processes and strategies that are irrelevant to those assumed by the test item should be as minimally utilized by test-takers as possible because they represent construct irrelevant factors. Accordingly, the research procedure proceeded in three steps: elicitation of expert judgment of the intended reading processes of the test, exploration of the test-takers' actual reading processes, and a comparison of the two.

Rationale for looking into test-takers' reading processes

Besides test method effects, another source of construct irrelevant variance, as claimed by Messick (1989), is test-wiseness. Test-wiseness occurs when test-takers capitalize on extraneous clues in item or test formats to achieve construct scores that are invalidly high. Test-wiseness strategies forms a constitutive component of the test-taking strategies taxonomy by Cohen & Macaro (2007) which also includes language learner strategies and test management strategies. However, unlike language learner strategies and test management strategies which may relate to part of learner ability under evaluation, test-wiseness strategies only look at peripheral aspects of the constructs being measured. Test-wiseness strategies in particular and test-taking strategies in general are believed to have engendered the practice of "teaching to the test" prevalent in many EFL contexts including Vietnam (Sadighi, Yamini, Bagheri, & Yarmohammadi, 2018). This practice directs students toward using tips, clues and test-wise knowledge to deal with particular test items for the purpose of achieving good scores on the test (Cohen, 2013).

Research into item difficulty and surface features of tasks only tells us how well readers have performed on the test in terms of their ultimate test scores while the question of whether readers draw on the actual processes intended to be elicited by the test or on test-wise strategies to produce that performance is left unresolved. In other words, until we know precisely what processes drive test-takers to choose the correct answers and whether these processes correspond to those intended by the test designers, score-based inferences about the test-takers' ability are deemed inconclusive.

Another concern that has been raised in the previous sections is the multi-process and multi-component nature of reading. A reader may arrive at a correct answer by employing a combination of various skills, processes, and strategies. These combined processes cannot be detected just by looking at the readers' test scores and should not be overlooked as it is highly possible that an item designed to test a particular skill or process turns out to be testing a number of different other skills and processes. Without a thorough inspection of this, important validity

implications may be missed. Therefore, in addressing research question 1, a significant amount of supplementary information will be yielded to support other research questions in the explanation inference as it delves into the internal processes of test item solving rather than looking at the test score only.

The data collection and analysis procedure for RQ1 is conducted via two primary stages: an expert judgment stage where reading processes intended to be measured by the test items were coded, and a stimulated recall stage where actual reading processes used by the test-takers to answer the test items were explored, each of which is discussed in the following sections. Figure 4.1. provides an illustration of the data analysis procedure for RQ1.

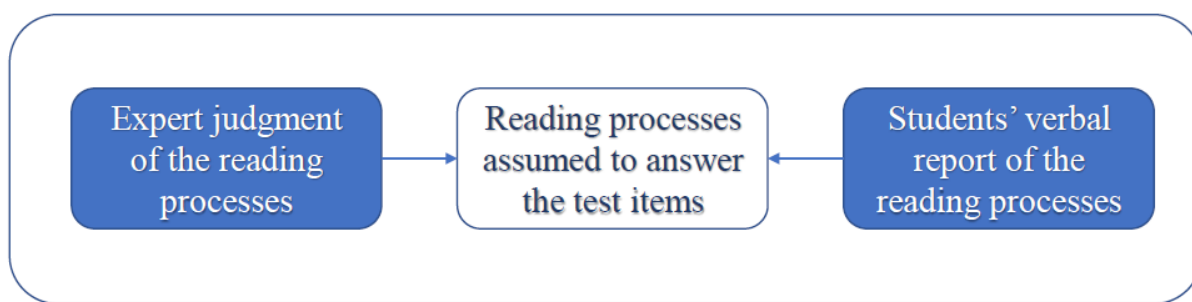


Figure 4. 1. Data analysis procedure for RQ1

Expert judgment

Description of the participants

Following the call for participation, two EFL lecturers (lecturer A and B), who met the recruitment requirements, gave their consent to participate in the study as experts to make judgment on the reading processes/skills required to answer the test items. They are both PhD candidates in Applied Linguistics who have earned Master's degrees in the same discipline at overseas universities and have finished courses in language testing and assessment as parts of their master programs. They are also EFL lecturers at UA where the L-VSTEP test was developed and administered. Lecturer A has ten years of experience teaching and assessing general English skills to undergraduate students in various disciplines, while lecturer B has nine years of experience teaching English reading and speaking skills to English, and non-English major students at the same university. In terms of reading assessment experience, they have both participated in training workshops for the L-VSTEP test item writers and assessors organized by the Ministry of Education and Training. They both are actively involved in the development, evaluation, administration, and

marking of the English reading tests for students across different disciplines at their own institution on a regular basis.

Instruments for data collection

The L-VSTEP test:

Practice form A of the L-VSTEP test was used in this phase of the study. The test is composed of four reading texts, each followed by 10 multiple choice questions (MCQ). The four reading passages had a combined total of 1,993 words, with Text 3 being the longest and Text 2 the shortest. The reading texts are arranged in order of difficulty, with Flesh Reading Ease measures for the texts being 75.5, 69.2, 53.7, and 44.2 respectively. Table 4.3 provides details of the Coh-matrix analysis of the four reading passages in terms of their length, readability, lexical features, syntactic complexity and referential cohesion.

Table 4. 3. Textual features of the L-VSTEP reading test Form A

Textual features	Text 1	Text 2	Text 3	Text 4
Text length (number of words)	513	464	527	489
<i>Syntactic complexity</i>				
Sentence length	13.865	16.571	22.913	21.261
Left embeddedness	2.676	2.25	6.609	6.217
Noun phrase density	0.605	0.949	0.992	0.736
<i>Lexical features</i>				
Type-token ratio	119.380	89.661	141.593	133.588
Word frequency	2.450	2.245	2.128	2.278
Word familiarity	582.080	573.261	556.135	562.477
<i>Cohesion</i>				
Referential cohesion (content word overlap)	0.044	0.032	0.021	0.067
Conceptual cohesion (LSA) - 40	0.093	0.132	0.107	0.151
Text readability	75.506	69.291	53.717	44.227

The expert judgment form:

Alderson and Lumley (1995) argued that the consistency of expert judgment of reading skills tested by an item can be seriously affected if there is no substantial agreement with respect to the interpretation of the skill descriptions. Therefore, Lumley suggested the construction of an

appropriate set of subskill descriptions and the achievement of consensus among judges as to the interpretation of these subskill descriptions. This two-step procedure was followed in the current study.

Skill descriptions employed in previous studies that involved human judgments of reading subskills were primarily adapted from existing skill lists and taxonomies (Alderson & Lukmani, 1989b; Anderson, 1990; Lumley, 1993). One limitation associated with this practice, however, is that the skill descriptions adopted might not be commensurate with the technical specifications of the reading test/task under investigation. In other words, unless the reading test/task is constructed to measure reading subskills as indicated in the test/task specifications, the application of alien taxonomies/skill lists in the judgment practice may not be appropriate (Alderson & Lumley, 1995). Therefore, the development of skill descriptions in this study took the guidelines for test item writers as a point of departure. Eight reading skills and relevant descriptions specified in the guidelines for test item writing were selected and placed in the expert judgment form, with each skill constituting a row, while the skill definitions and descriptions laid out as columns. Since the guidelines were originally developed in Vietnamese, the skill definitions and descriptions were translated into English by the researcher for the purpose of reporting. The expert judgment form was then subject to an extensive discussion session where consensus among experts in terms of the interpretation of the skill descriptions was sought. Details of this phase are presented in the next section.

Procedure for expert judgment

It is well-documented in the literature on reading assessment that the process of responding to a test item may involve the use of more subskills than what was originally intended by the test constructor (Alderson & Lumley, 1995; Dawadi & Shrestha, 2018; Tengberg, 2018). By disregarding this nature of reading processing and applying a superficial skill-matching approach to the expert judgment process, important validity evidence of the reading assessment may be overlooked, and the judgment outcome may introduce bias. As a result, the experts in this study were required to identify not only the primary reading skill targeted by each item, but also the potential involvement of other skills in responding to a particular item.

Each of the experts was required to identify the primary reading skill/process intended to be activated by each test item as well as the mutually supplementary processes they believed were involved in answering the item. The experts' judgment was informed by the guideline for test item

writing that was used for developing the test at UA. For other reading processes that match none of the processes detailed in the guideline, experts were asked to write a brief description of the processes and provide their own judgment based on the description.

Initially, the experts were invited to participate in a training session. This training session entails two purposes: to familiarize the experts with the guideline and judgment protocols. As such, essential information about the L-VSTEP reading test, as well as concepts and categories of the guidelines were clearly explained to the experts. They then were required to take a sample version of the L-VSTEP reading test (one reading passage with ten items). Each individual expert judged the test items by ticking in the relevant box that indicates the primary reading skill/process as well as the supplementary reading skills/processes they believe are tapped by the item. In addition, experts were encouraged to report any judgment categories that were unclear or confusing which were then discussed and reconciled in a subsequent moderation session.

The main expert judgment session followed exactly the same procedure as the training session, but with the official version of the L-VSTEP reading test (version A) used in the study. The whole judgment process was conducted independently by the experts who were allowed to take the testing and judgment materials home so that they had as much time as they needed to deliver reliable and accurate judgment of the reading processes. A face-to-face discussion was organized afterwards to make sure that any disagreements or discrepancies were resolved before the final judgment was established. This is an important step in the quality assurance process which is normally followed to ascertain the reliability of the judgment outcomes. The final judgment protocol was used as a proxy for the subsequent comparison with learners' actual reading processes.

The data analysis was conducted on an item basis where the skill(s) jointly agreed by the experts to be tested by the reading test items were closely examined, and emerging patterns of skill involvement across the items were then reported. Since a) the experts worked on the same expert judgment form which was derived from the test development guidelines, b) consensus as to the interpretation of skill descriptions has been established, and c) both of them participated in a moderation session during which they discussed to achieve agreement on the judgment of skills, it was decided that there was no need to resort to quantitative measures of inter-rater reliability.

Verbal report methodology

An essential step that must be taken to address the research question is to elicit information about how test-takers go about answering each of the reading test items. This can be achieved via the introspective verbal report method (Gass & Mackey, 2017) which garners data by asking participants to vocalize what goes through their mind as they are performing a specific task or activity. Since there are different categories of verbal report depending on the question asked, the nature of information collected and the procedure for collecting data, initial description and distinction of the different verbal report methods is important to clarify as it helps to remove any methodological ambiguities.

Gass and Mackey (2017) distinguish between two types of report - think aloud protocol and stimulated verbal recall according to three categories: time frame, forms of report, and support types. Think aloud protocol is data collected in oral form while the learners are working on the task with the only support being the event. On the other hand, stimulated verbal recall refers to post-event data documented either in written or oral forms, and with different types of support such as video or audio recording and observation notes. It is this latter verbal report type that constitutes the predominant method of data collection used to address research question 2. In the following section, the “why”, the “what”, and the “how” – that means the reasons for, the fundamental issues associated with, and the procedure for collecting data using stimulated verbal recall method - are delineated.

Rationales for the use of stimulated verbal recall

Stimulated verbal recall is used to elicit learners’ thought processes after the task completion. It is preferable over the think aloud protocol in the current study for three reasons. First, think aloud protocol may alter the naturally-occurring cognitive processes when one engages in a reading activity, thus rendering the verbalized reports less reliable (Gass & Mackey, 2017; Green, 1998). This holds true particularly for the current study because test-takers are likely to suffer from splitting attentional resources as they have to perform different cognitively engaging activities within the context of a standardized test. Second, think aloud protocol leads test-takers to using more time than normal to complete the test. Fatigue or boredom may ensue from this practice, which seriously affects task completion. Furthermore, without a standard time limit as in a real standardized reading test, the distinction between lower and higher levels of cognitive processes as a corollary of time pressure (see chapter II for a discussion of this) may be obscured. Finally, as pointed out by Bloom (1954), think aloud protocol involves intensive training of

participants, the success of which may not always be as expected because “even after training, not all participants are capable of carrying out a task and simultaneously taking about the task” (p.23). Stimulated verbal recall is also preferable over self-observation (Cohen, 1984) in that the latter draws heavily on memory without any prompts (Bloom, 1954, p.26). Stimulated verbal recall also involves memory retrieval, but with the support of different types of stimuli.

Caveats to using stimulated verbal recall

Since data elicited as part of the stimulated recall procedure is purely qualitative, the issue of validity and reliability of the procedure merits careful considerations.

Validity is primarily concerned with the extent to which what test-takers report corresponds with what they actually think when they engage in the tasks. Validity of the method is likely to be compromised when incomplete reports, distorted or additional information, and disruption of the test-taking process occur frequently (Cohen, 2007; Gass & Mackey, 2017; Green, 1998; Nisbett & Wilson, 1977). Several measures can be undertaken to uphold the validity of the method. For example, training sessions with appropriate instruction on how to implement the reporting procedure should be offered to participants to make sure that they are aware of what they should and should not report. In addition, the time lapse between the test event and the recall interview should be kept to minimum because the longer the time lapse, the more likely that the reporting practice is susceptible to memory loss. Language of recall should also be considered as it may affect validity of the method. It is recommended that participants use either the target language or their mother tongue to report on their thought processes, whichever they are comfortable with.

Reliability refers to the consistency of report data given similar participants and similar tasks. Reliability of the method can be improved by carefully training the verbal protocollers and standardizing the data elicitation procedure. Individual difference, such as language proficiency, can affect reliability because learners with differing proficiency levels may approach the task differently and use different strategies and processes from each other. This, on the one hand, attenuates the consistency of coding measures, but on the other hand, provides essential information about cognitive processes as they are employed by learners of different proficiency levels. It is, therefore, necessary to categorize learners into groups of different proficiency levels and explore how different their cognitive processes are. Task variability may also present another source of reliability threat as different tasks may induce different cognitive processes. This is, however, of little concern in the current study because all L-VSTEP reading test items are in

multiple choice format, and variability in cognitive processes, if any, is not an artifact of the task facet. Being aware of the validity and reliability concerns in using stimulated verbal recall, the following section summarizes the procedures for conducting data collection and analysis using this method.

Procedure for collecting stimulated verbal recall data

As discussed in section 4.3.2, major concerns regarding the collection and analysis of stimulated verbal recall data are their validity and reliability. The former concerns the quality and accuracy of the test-takers' verbal reports in relation to their thought processes, while the latter refers to the consistency with which recall data from different readers can be elicited. A major threat to the validity of stimulated verbal recall is non-veridicality which indicates the inefficiency of verbal reports due to omission or commission of thought processes (Russo, Johnson, & Stephens, 1989) during reporting. In order to alleviate potential threats to the validity and reliability of the stimulated verbal recall data, careful consideration should be given to both the data collection and data analysis procedures. Details of these procedures are illustrated in Figure 4.2 following Green (1998) suggestions for collecting and analysing stimulated verbal recall data.



Figure 4. 2. Procedure for collecting and analysing stimulated verbal recall data

Instruments

Instruments for the stimulated verbal recall study include the L-VSTEP test version A, a sample task, guideline sheet and the written consent form. The L-VSTEP test was used both as a stimulus for the verbal report and in the expert judgment phase. Detailed description of the test was provided in section 5.2.2. The sample task was taken from a L-VSTEP practice test and was used during the training session. Detailed instructions on the procedure to conduct stimulated verbal recall and important caveats to be considered when reporting were included in the guideline sheet which was accompanied by the researcher's verbal instruction during the training session.

Description of Participants

In order to gain insights into the reading processes of the test-takers, a group of nine students were invited to participate in this phase of the study. One criterion for selecting the participants was that they should be selected members of the targeted population of the L-VSTEP reading test – they were students at UA where the test was used and were planning to sit the test

as part of the English proficiency requirements for graduation. Another criterion was that they should be selected from three distinct groups of English reading proficiency as indicated by their scores on the test. The level three group includes three participants who scored lower than 5.5 on the test. Three participants in the level four group had scores in the range of 6 – 8, and the other three in the level five group scored higher than 8 on a 10-point scale. The classification of participants according to proficiency levels allows for an insightful examination of the reading processes employed by students with different proficiency levels, thereby offering useful evidence about the test’s discriminatory power as well as the extent to which it elicits readers’ reading processes along a proficiency continuum. Table 4.4 provides background information about the verbal protocol participants.

Table 4. 4. Background information of the participants

Codes	Gender	Age	Time spent learning English	Scores on the test	Levels of English reading proficiency as indicated by their scores
Student 1	M	24	13 years	9	5
Student 2	F	21	10 years	9.5	5
Student 3	F	22	11 years	9	5
Student 4	F	21	10 years	7	4
Student 5	F	22	11 years	7.5	4
Student 6	M	23	12 years	6	4
Student 7	M	22	11 years	4.5	3
Student 8	M	22	11 years	4	3
Student 9	F	21	10 years	4	3

Training

The nine participants took part in a training session. The purpose of the training session was to familiarise them with the purpose of the study, the format and requirements of the L-VSTEP reading test, the procedure for conducting the verbal report, and first-hand experience with a trial verbal report session. Accordingly, after giving their consent to participate in the research, participants were informed about the study purposes, essential information about the test, and steps

to follow when doing stimulated verbal recall. An important clarification was that participants were required to report their reading processes rather than to explain their answers to the test items as the latter might induce non-veridical reports.

Participants were first shown a video about a stimulated verbal recall session, the purpose of which was to give them a sense of how the session was conducted. After watching the video, participants were given a short sample of the L-VSTEP reading test comparable to the one used in the official session in terms of text length, format, and difficulty level. Immediately after the test, participants reported on their reading processes while answering each item, following both the verbal and written instructions directed to them earlier. The whole process was under the instruction and observation of the researcher to make sure that essential information was clearly articulated and understood by the participants. The training session concluded with some comments and feedback by the researchers so that participants were aware of what was required and what to avoid (Green, 1998), as well as participants' opinions about the factors that might affect their verbal report.

The training session took place one week before the main data collection stage and took each participant approximately one hour and a half, including 10 minutes for video viewing, 30 minutes for completing the sample test, 30 minutes for verbal report and 20 minutes for break and downtime. The total amount of time for the training session was noted by the researcher to facilitate the preparation of the procedure and the estimation of the time for the main stimulated recall data collection phase.

Data collection

The main data collection session was conducted with each individual participant in a quiet room and was audio-recorded. The procedure for this stage was the same as in the training session with some caveats informed by participants during training. First, the language of reporting was Vietnamese as it is the mother tongue of all the participants, and it made the verbalization easier. Second, several participants expressed their concerns about the presence of the researcher while they were reporting, which raised concerns about the veridicality of their reports. Therefore, where it was not necessary, the researcher stood at a distance while some students reported and minimal interruption was made during the reporting session. Finally, as observed in the training session, the disclosure of answer keys unexpectedly altered the recall of reading processes and induced

participants to produce reactive responses. As a result, answer keys were not given prior to the main stimulated verbal recall.

Data analysis

The nine verbal reports were transcribed verbatim by the researcher, and only relevant segments that were used for thesis supervision purpose and quoted in the thesis were translated into English. The transcripts were organized on an item basis resulting in 360 protocols, each representing a participant's responses on one item. Units for analysis – a phrase, clause, or a sentence - were identified for each protocol prior to the segmentation process. The segmentation was conducted with the basic idea that each segment represents a single process or strategy (Green, 1998).

The segments were analysed thematically and went through several rounds of coding and recoding. First, via the process of listening to the recordings repeatedly, transcribing and segmenting the protocols, the researcher has developed familiarity with the content of the protocols and visualized how to code the segments as well as built up some initial codes in mind. As the coding progressed, new codes emerged from the data that reflected the nature of the test-taking process. For example, some unique strategies that each student employed to answer the test items were identified beyond those specified in the test development guidelines. Six additional codes were identified, namely Eliminating Implausible Answers (EIA), Keyword Matching (KM), Test Taking Experience (TTE), Replacing For Confirmation (RFC), Uninformed Guessing (UG), and Syntactic Parsing (SP). In addition, the subskill Lexical Inferencing was further refined to reflect the actual deployment of this subskill in the reading process since protocollers exercised not only this skill but also their lexical knowledge to decipher the meaning of unfamiliar words. As such, the process of lexical access was also added to the list of reading processes.

Intercoder reliability was then conducted to make sure that the researcher was not heavily influenced by his study hypotheses and expectations, and thus the coding was free from idiosyncrasy (Green, 1998). Twenty percent of the 759 segments was co-coded by an independent coder who was also a PhD candidate in language testing and assessment and was familiar with the L-VSTEP test. These segments represented the whole spectrum of reading subskills identified in the expert judgment phase and therefore, lent themselves to further triangulation from independent coders. The percent agreement was calculated simply by dividing the total number of codes agreed by both coders by the total number of co-coded segments. The coders reached agreement on 92

percent of the co-coded segments, which supported the intercoder reliability of the study. Where disagreement occurred as to which subskills the segments should be assigned to, the two coders worked together and discussed alternatives until final agreement was achieved.

Finally, the data analysis involved the comparison of the coding protocols pertinent to the test-takers and those of the experts. Comparison was made on whether the expected reading processes of each test item corresponded with the reported processes of the test-takers, whether there were additional processes used by the test takers and how they contradict, supplement or extend on the intended processes, whether test-takers at different proficiency levels employed different processes, and how these different processes, if any, were related to the expected processes.

4.3.2. Research question 2

To what extent is the factor structure of the L-VSTEP reading test consistent with a proposed theoretical model of the test construct? Is the factor structure of the test invariant across groups of test-takers with different reading proficiency levels and academic disciplines?

This research question responds to the most abstract plane of Chapelle et al.' (2008) model of explanation inference, the major focus of which is to ascertain that the internal structure of a test reflects highly interrelated components with respect to theoretical expectations. In other words, the research question aims to examine the extent to which the factor structure of the L-VSTEP reading test's scores conforms to the underlying theoretical model of the test construct and whether this factor structure remains statistically consistent across groups of test-takers with different reading proficiency levels and academic disciplines. In order to address the research question, two assumptions are formulated to underlie the research design stage. First, the underlying structure of the L-VSTEP reading test scores should comply with the test's underlying theoretical construct of general reading proficiency as proposed by the test designers. This underlying theoretical construct is illustrated in the guideline for test item writing (see chapter I) as consisting of a set of skills/subskills that inform the test design process. Second, if the factor structure of the test scores proves identifiable and justifiable through statistical modelling procedure, that factor structure should remain invariant across groups of test-takers with different reading proficiency levels and academic disciplines. The rationales which inform these assumptions and which in turn, are informed by relevant literature as well as the procedure for data collection and analysis to answer research question 2 are discussed in the following sections.

Rationale for examining the factor structure of the L-VSTEP reading test

The questions of whether reading subskills exist or what patterns underlie them have been empirically tested by numerous researchers in both L1 and L2 contexts (see chapter II). However, various factors including, but not limited to, readers proficiency, task characteristics, research methods, classification schemes and the very test used in each study have combined to make a definitive answer elusive. This presses the need for developers of any standardized reading tests, including the L-VSTEP test, that claim to measure a certain reading skills or subskills among the readers to prove that their test actually does so in a reliable and appropriate manner, and with theoretical as well as empirical justifications. In consideration of the studies reviewed in chapter II, there are three important factors that warrant further empirical examination in this phase of the study: the readers' proficiency and relevant academic disciplines, the classification schemes, and the research methods used.

The influence of the reader language proficiency factor on whether or not L2 reading skill is divisible has been a matter of debate among previous scholars. For example, Carr and Levy (1990), Alderson (2000), and Song (2008) argued that L2 reading subskills are more readily identifiable for beginning, weak or low-level second language readers than for advanced readers. This is partly because of the assumption that advanced readers have achieved automaticity in lower-level processes while higher-level processes and associated reading skills become so integrated that it is hard for them to be separated. However, this position is at odds with Van Steensel, Oostdam, and Van Gelderen (2013) who found no divisible patterns among the test items in both high- and low-level students. This indeterminacy of the effect of language proficiency on the divisibility of reading comprehension should be given due consideration if the nature of reading subskills is to be understood more clearly. Surprisingly enough, there have been no empirical studies that take language proficiency into account in uncovering the underlying patterns or describable subskills of reading tests. This gap should be addressed in the current study because a reading test that measures certain skills for a given group of readers (e.g. high-level students) but different skills for another group (e.g. low-level students) may negatively influence the interpretation and use of the test scores.

Another individual characteristic factor that seems to be missing from previous studies on L2 reading skill divisibility is the effects of reading experience in a specific language use context. The general literature in L2 reading has documented the roles of experience with reading exposure

in learners' reading abilities as a potential difference between L1 and L2 reading (Grabe, 2009). However, little empirical evidence is available to shed light on whether experience with reading exposure among students with different academic disciplines in an academic context has any effects on the (in)divisibility of their L2 reading ability. This gap is also addressed in the current study.

Different classification schemes have been used to categorize subskills. For example, Sawaki et al. (2009) used reader purposes as detailed in the TOEFL iBT test specification to propose three subskills: basic comprehension, reading to learn and inferencing. Van Steensel et al. (2013) proposed three subskills indicating three levels of understanding: retrieving, interpreting and reflecting. Song (2008) hypothesized three aspects of reading: understanding main/topical ideas; understanding supporting/specific details; and making inferences. Kim (2009) tested three types of reading: reading for literal meaning, reading for implied meaning within the texts, and reading for implied meaning beyond the texts. Tengberg (2018) used the reading process categories stipulated in the Swedish National Agency for Education for classification: retrieve explicitly stated information; make straightforward inferences; integrate and interpret information and ideas, and reflect; and examine and evaluate content, language, and textual elements.

A more careful look at these classification schemes reveals that many are based on conceptual and theoretical considerations while few have aligned the proposed skills directly to the constructs defined for the very test being used and operationalized in the test specifications. It should be noted that test items are designed based on test specifications, which in turn are derived from the constructs defined for the test. Any effort to uncover the underlying patterns of the subskills of a specific test without considering how these skills are described by the test designers may provide misleading or distorted information. This study, therefore, takes the reading subskills described in the test development guidelines as the basis for proposing the hypothesized underlying structure of the L-VSTEP reading test.

Different research methods, both qualitative and quantitative, have been employed to discern the (in)divisibility of reading comprehension. One approach is to use exploratory factor analysis to statistically examine if putatively different variables (e.g. subskills) function in a similar manner. If all putative subskills load on a single factor, reading is considered a unitary concept. On the other hand, if some putative subskills load heavily on one factor while other subskills load on other factors, this suggests that reading is multi-divisible (Weir & Porter, 1994).

Another approach is to use confirmatory factor analysis to testify a priori hypothesized underlying patterns of reading comprehension. If an adequate fit between the hypothesized models and the data collected was achieved on the basis of substantive and statistical considerations, the subskills are determined according to the hypothesized patterns.

A number of other studies take a somewhat more qualitative approach. As such, several reading subskills are proposed to explain item processing in a reading measure. Experts (e.g. teachers, researchers) are then required to judge and match those subskills with the specific items independently of each others. If they can reach a substantively sufficient agreement in their classification, then reading is considered divisible. Conversely, lack of agreement among raters indicates that no divisible subskills can be extracted. However, this approach has been called into question by other scholars (Alderson, 2000; Weir & Porter, 1994). Alderson (2000) argued that the focus of any tests should be on the test-takers, so what matters is not what the judges or raters think an item is testing, but rather the actual processes engaged by the readers. The thought processes of raters may be fundamentally different from each other and from those of the readers, which is likely to lead to inconsistency among raters and between raters and readers. Furthermore, lack of training, lack of shared understanding of the targeted skills and inconsistency in skill descriptions (Tengberg, 2018; Weir & Porter, 1994) also prove major threats to the validity of this methodological approach. These considerations prove the need to bring together both quantitative and qualitative approaches in a mixed paradigm for triangulation purposes.

In this phase of the study, both expert judgment and confirmatory factor analysis (CFA) were employed to answer RQ2. First, participants are required to take version A of the L-VSTEP reading test within the given standard time (60 minutes), which produces samples of test scores for CFA and measurement invariance. At the same time, the expert judgment data yielded in the verbal report stage are used to inform the specification of the CFA models in this stage. The judgment data were focused on the assignment of the test items in version A of the test to the corresponding reading skills as specified in the test item writing guidelines. Exploratory factor analysis was not used because the validation paradigm proceeds in a confirmatory mode anchored in a well-defined theoretical construct illustrated in the guideline for test item writing rather than in an exploratory fashion where the construct is the subject of exploration. The research process, therefore, consists of two steps: Confirmatory factor analysis of the hypothesized factor structure

of the test and the examination of the measurement invariance of the identified factor structure (See Figure 4.3).

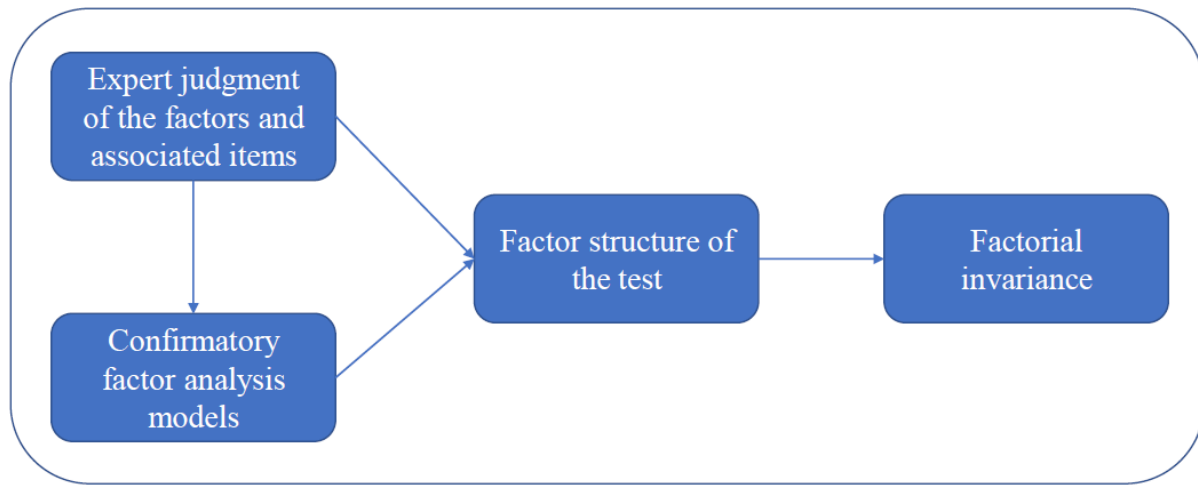


Figure 4. 3. Data analysis procedure for RQ₂

Participants

Following the call for research participation, 544 undergraduate students out of approximately 4000 last-year undergraduate students across different academic disciplines at UA gave their consent to participate in the research and completed the test form A on the scheduled dates. All participants are final year students pursuing their undergraduate studies in various disciplines at UA where the L-VSTEP is used as an English proficiency screening measure for graduation. Prior to giving consent to participate in the study, all participants indicated that they were familiar with the format of the test and are either taking test preparation courses offered by the English centre at the same institution or doing practice tests at home. Regarding the participants' demographic information, there is a large disproportion in terms of gender. Only 62 (11.4%) participants are males while 482 (88.6%) are females. At the time of data collection, 26.3% (N = 143) of the participants were doing their bachelor's degree in English teacher education, 39% (N = 212) in English for translation and interpretation, and 34.7% (N = 189) in non-English majors. As indicated by their performance on the test, 6.5% (N = 35) of the participants are at level 5 according to the CEFR-VN which is comparable with level C1 of the Common European Framework of Reference (CEFR), 68.6% (N = 369) at level 4 (B2), and 25.7% (N = 140) at level 3 (B1). They ranged in age between 20 and 22 years (M = 21.17, SD = 0.81),

with six participants missing a value on age. Applying the general rule of thumb for calculating sample size for Confirmatory Factor Analysis (CFA) as well as following the suggestions by Cohen (1992) and Bentler and Chou (1987), the cohort of 544 participants are sufficient for both the testing of the CFA models and the measurement invariance analysis in the current study.

Confirmatory factor analysis

The primary purpose of the research question is to relate the underlying pattern of the L-VSTEP reading test scores to a hypothesized factor structure of the test as informed by theoretical considerations pertaining to the test construct. Theoretical underpinnings that underlie the hypothesized factor structure were derived from relevant literature in the field and the L-VSTEP reading test guideline for test item writing. A review of literature on the notion of subskills in second language reading (see chapter II) highlights the fact that L2 reading is a complex concept. There seems to be a considerable divergence among L2 reading researchers as to whether L2 reading comprehension is a unitary or divisible concept; and if it is divisible, how many skills/subskills can be discerned. This situation applies to the L-VSTEP reading test because the test guideline operationalizes the construct of general L2 reading proficiency according to the types of reading subskills that serves as the guiding principles for the item development and validation. The subskills include understanding explicit information, understanding cohesive devices, lexical inferencing, understanding pragmatic meaning, inferring situational meaning, integrating information, and summarizing information (MOET, 2015b; Nguyen, 2018). These subskills serve as the theoretical “underlying factors” in the building of the competing hypothesized models of the L-VSTEP reading test structure reflecting competing theories about L2 reading comprehension.

In consideration of an appropriate statistical procedure to testify the hypothesized factor structure of the L-VSTEP reading test, Confirmatory Factor Analysis (CFA) emerges as a prime candidate. CFA enables a simultaneous analysis of the interrelationship among the entire system of variables to examine the consistency between the hypothesized model and the collected data via the estimation and evaluation of the goodness-of-fit indices (Byrne, 2010). Contrary to multiple regression analysis or path analysis which only deal with observable or measurable variables, and which assume that all variables are measured without errors, CFA incorporates both observed and unobserved (or latent) variables into the modelling process. This is a distinct advantage of CFA because numerous variables in human science represent hypothetical constructs, just as the

constructs of L2 reading proficiency or L2 reading skills/subskills in the current study, that can neither be directly observed nor perfectly measured, and therefore must be indirectly estimated from associated observed variables (Byrne, 2010; Hair et al., 2014; Wang & Wang, 2012). It is the incorporation of the latent variables in CFA that allows for the estimation of, and control over, the measurement errors, thereby considerably reducing the statistical modelling inaccuracies and biases vis-à-vis the prevalence of these artifacts in other error-free measurement methods. A standard procedure for conducting CFA usually proceeds in five steps: model specification, model identification, model estimation, model evaluation, and model modification. The following paragraphs discuss these steps in turn as well as their relevancy to the current study.

Model specification

Model specification involves the construction of a hypothesized theoretical model based on relevant theories and related empirical research. The goal of model specification is to determine the variables to be included in the model, the relationship among these variables, and the parameters to be estimated such that the model implied variance-covariance matrix sufficiently approximates the sample derived variance-covariance matrix. In other words, the hypothesized model as informed by relevant theoretical considerations should be precise to the extent that the discrepancy between that model and the true model derived from the sample data is not sufficiently large so as to lead the researcher to rejecting his/her hypothesized model. Three issues that bear significant importance to the specification of the hypothesized model should be given due consideration, namely the identification of exogenous variables and endogenous variables, the importance of theories in model building and the selection of modelling strategy.

Two types of variables exist within a CFA model, endogenous variables and exogenous variables. Endogenous variables are those whose variability is theoretically determined by other variables within the CFA model, whereas exogenous variables are influenced by unknown variables outside the CFA model (O'Rourke, Hatcher, & Stepanski, 2005). For example, in the CFA model to be established in the current study, endogenous variables are the underlying factors of the L-VSTEP test while the exogenous variables are the indicators or items that load onto their hypothesized corresponding factors.

The specification of any CFA models must be theory-driven. This means that the identification of variables, their relationship, and the parameters to be estimated in a CFA model cannot grow out of atheoretical concerns, but instead must be dictated by relevant theoretical and

empirical justifications and must be substantially meaningful (Ho, 2014). In the current study, the theoretical foundation for building the CFA model manifests itself in the extensive review of extant literature and in the guidelines for test item writing that guide the test item development process.

Modelling strategy involves the strategic approaches to the execution of the CFA procedure (Hair et al., 2014). Confirmatory modelling strategy is a straightforward approach where a single model is specified and then adopted or rejected on the basis of goodness-of-fit indices. However, the potential rejection of the model based on goodness-of-fit indices may extravagantly dismiss the significant amount of time and resources invested in the building and testing of the model. Model development strategy may partly address the above problem by adopting a modification strategy whereby no straight rejection of the model is made given inadequate fit. Instead, the model is modified until it achieves good fit. Finally, the model competing strategy involves the specification of several alternative models. In so doing, the rejection or adoption of the ultimate model is based on the relative goodness of fit of the models, the adequacy of the parameter estimates, and model parsimony. The latter two strategies were adopted in this study.

Model identification

Model identification refers to the capability of the model implied variance-covariance matrix to produce a unique set of parameter estimates given the observed data contained in the sample derived variance-covariance matrix (Schumacker & Lomax, 2010; Wang & Wang, 2012). In other words, it is essential to determine in advance the number of data points available and the number of parameters to be estimated so that a unique solution to each and every equation expressed as a function of the hypothesized model can be achieved. The difference between the number of available data points and the number of free parameters to be estimated constitutes the model's degree of freedom. A model is considered over-identified if its degree of freedom takes on positive values – that means there are more data points available than free parameters to be estimated. If degree of freedom is negative, the model is under-identified and if degree of freedom equals to zero, the model is just-identified. Only over-identified model is of concern in the current study because just-identified models test no theory while under-identified models cannot be estimated. Another point of concern is that latent variables in the specified model have no measurement scale, resulting in the scale indeterminacy problem (Byrne, 2010; Schumacker & Lomax, 2010). To overcome this problem, parameter constraints must be imposed (fixed to the value of 1) either on the variance of the corresponding latent variable or on one of its factor

loadings. Applying these rules to the models tested in the current study, it can be determined that the identification requirement has been satisfied in all three models.

Model estimation

Model estimation involves the employment of a particular fitting function so that the difference between the model implied variance-covariance matrix and the sample derived variance-covariance matrix is minimal (Byrne, 2010; Schumacker & Lomax, 2010). The most popular CFA estimation method is the Maximum Likelihood (ML) estimation which is the default option in many CFA/SEM software programs. The use of ML method rests on several essential assumptions regarding the input data. One assumption is that variables should be measured on an interval/ratio scale. Recall that the VSTEP reading test is composed of only dichotomously scored items, this assumption is violated. One way to address this problem is to use item parceling technique which calculates the sums of responses to groups of intercorrelated items that load on the same factor and uses these sums in the subsequent latent variable analysis (Ho, 2014).

Another assumption of the ML estimation is that data must have multivariate normal distribution (Byrne, 2010). Therefore, univariate and multivariate distribution of data must be checked prior to the running of CFA using ML method. Univariate distribution can be examined by the inspection of skewness or kurtosis values – skewness values lower than 3 and kurtosis values lower than 10 (Kline, 2016). The necessary but insufficient condition of univariate normality should be followed by the inspection of the multivariate normality assumption. The Mardia's normalized estimate of multivariate kurtosis can be used to examine this assumption. As such, the Mardia's values higher than 3.00 are indicative of multivariate nonnormal data (Yuan, Marshall, & Bentler, 2002). Data should also be devoid of multivariate outliers. Multivariate outliers can be detected by the computation of the squared Mahalanobis distance (D^2) which represents the distance in standard deviation between scores for one case and the sample data mean (Byrne, 2010). D^2 for one case should not be in distinct distance from D^2 for other cases so as to rule out the existence of extreme scores that may affect the distributional assumption of data. In case of data that have multivariate nonnormal distribution, Satorra and Bentler (1994) proposed the use of a scaling correction for the χ^2 statistic (hence the $S - B\chi^2$) that takes into account the model, the estimation method, and the sample kurtosis values.

Model evaluation

Model evaluation focuses on determining how well the hypothesized model accounts for the observed sample data based on the evaluation of the global model fit and the appropriateness of the individual parameters.

The global model fit is assessed on the basis of numerous goodness-of-fit indices generated as a result of the modelling process. Adhering to the common reporting practice of SEM/CFA in the literature (Bentler, 2008; Brown, 2006; In'nami & Koizumi, 2011; Kline, 2016), and particularly to the suggestions of Ockey and Choi (2015) which are of direct relevance to second language testing contexts, five goodness-of-fit indices are reported in the current study: two absolute fit indices (the χ^2 statistic and the standardized root mean squared residual (SRMR)), one adjusted for parsimony index (the root mean squared error of approximation (RMSEA) and its confidence interval), one relative fit index (the Tucker-Lewis index), and the Akaike's Information Criterion (AIC). Although no absolute values or guidelines are available for the interpretation and determination of an acceptable model fit, several researchers, particularly Hu and Bentler (1999) have proposed cutoff values for these goodness-of-fit indices based upon simulation studies, which now become widely used in the literature (Ockey & Choi, 2015).

The conventional χ^2 statistic tests the hypothesis that the specified model is consistent with the observed sample data (O'Rourke et al., 2005). A small value of χ^2 with nonsignificant p value is indicative of reasonable model fit. However, due to its sensitivity to sample size and its assumption of perfect model fit, the χ^2 value should be augmented by other indices. The SRMR represents the standardized residual of the model implied and sample derived variance-covariance fitting process. The smaller the value, the better the fit. The cutoff value of .08 is suggested by Hair et al. (2014) for good model fit. RMSEA preserves model parsimony by penalizing models with more free parameters. The lower the RMSEA values, the more parsimonious the models. RMSEA values of .07 or lower with the 90% confidence interval in the range of $.00 < CL_{90} < .09$ are suggestive of acceptable model fit (Hair et al., 2014; MacCallum, Browne, & Sugawara, 1996). Tucker – Lewis Index (TLI) is an incremental fit index that compares the hypothesized model with a baseline model. TLI value ranges from zero to 1 with values higher than .90 indicating good model fit (Hair et al., 2014). Finally, the AIC and its consistent version (CAIC) (Bozdogan, 1987) address the issue of model parsimony in model comparison by taking into account the statistical goodness-of-fit, the sample size and the number of estimated parameters. Models with lower values of CAIC/AIC represents better fit (Hu & Bentler, 1999).

The assessment of the individual parameters in a model, such as variance, covariance, and factor loadings, hinges on three features. First, parameters should be significantly different from zero. This can be checked by the inspection of the critical value ($t > 1.96$ at $\alpha < .05$ level) which is calculated by dividing each parameter estimate by its standard error. Second, the size (values) and signs (positive or negative) of the parameters should conform to the hypothesized model's expectations and should be meaningful. This is also tantamount to the absence of Heywood cases (Byrne, 2010; Hair et al., 2014; Kline, 2016) which are characterized as correlation > 1.00 or negative variance (including error variance).

Model modification

Model modification involves the detection of diagnostic cues that represent the greatest sources of model misfitting and the reestimation of the hypothesized models that take those cues into account in an attempt to achieve greater model fit. These diagnostic cues are usually related to the standardized residuals produced as an artifact of the model fitting process and fixed parameters – parameters that are not freely estimated in the original model. Standardized residuals greater than 4.0 flag problematic variables while values in the range of 2.0 to 4.0 call for further attention and consideration (Hair et al., 2014). Modification indices can be consulted to identify originally fixed parameters that can be made free to achieve better fit in the reestimation of the model. Modification indices of 4.0 or greater are prime candidates for inclusion in the modification process (Hair et al., 2014). However, the temptation to achieve better model fit based on statistical criteria of modification indices should not override theoretical and substantive considerations. After all, informed theories and the meaningfulness of the relationship among variables should play key roles in both the specification and respecification of the model no matter how tempting the model fit may sound.

Factorial invariance of the L-VSTEP reading test

After the factor structure of the L-VSTEP reading test has been testified and confirmed by the sample data, the next step is to subject this factor structure to a factorial invariance analysis. This serves two purposes: to cross-validate the factor structure by imposing it on the observed data derived from different subsamples of the same sample (less proficient readers sample and more proficient readers sample; English pedagogy, English translation, and Non-English major samples), and more importantly to testify the hypothesis posed by previous researchers that the divisibility of L2 reading subskills is not transparent with regard to readers of different proficiency

levels and that the divisibility of L2 reading subskills may be subject to reading experience in different academic disciplines. This will also reveal potential implications for the interpretation of test scores because, as pointed out earlier, a reading test item that measures different subskills across different targeted samples may weaken the interpretation and use of the test scores.

Factorial invariance, or factorial equivalence, refers to the test of invariance of the underlying structure of a measurement instrument to see if items in the instrument behave in an identical manner across multiple samples, and if not, where the sources of noninvariance can be traced (Byrne, 2010). Factorial invariance testing normally proceeds in a hierarchical sequence during which results of the former steps determine what needs to be done in the subsequent steps. As preliminary steps to the testing of factorial invariance, group membership must be determined and a baseline model must be established. For the current study, two subsamples of reading achievement and three subsamples of academic disciplines were created. The former is based on test-takers' scores on the L-VSTEP reading test in the main study using the mean value as the cut-off value. An independent sample t-test is conducted afterwards to ascertain that the two groups are statistically different in terms of their reading proficiency. The latter is based on the academic programs which the participants are following at the relevant institution.

The baseline model is established by imposing the CFA-derived factor model on data from the different groups of test-takers separately. Goodness-of-fit evaluation and model modification are employed to arrive at a single baseline model that is identically specified for both groups. This baseline model is then tested in the initial step of the main invariance testing stage by simultaneously incorporating data from both groups into a single analysis. Result of this analysis is a configural model whose goodness-of-fit values serve as the baseline values against which the subsequent measurement invariance models are compared (Byrne, 2010). Since the primary purpose of the current study is to examine if the loading patterns of the test items remain consistent across samples, only metric invariance was tested. Metric invariance was tested by imposing equality constraints on all factor loadings of the configural model and comparing the configural model with the metric model by consulting the values of the χ^2 difference test. A non significant value of the χ^2 difference test should be expected if measurement invariance is to be achieved, indicating the model to be statistically invariant across samples of test takers with different levels of English proficiency in reading and different academic disciplines. Failure to yield nonsignificant value of the χ^2 difference test indicates measurement noninvariance which, in turn, necessitates

the search for problematic items. This can be accomplished by assigning increasingly restrictive constraints to the loadings at item levels respectively until the problematic items are identified.

4.3.3. Research question 3

What are the linguistic and discourse characteristics of the texts, items and item-text variables of the L-VSTEP reading test? To what extent do they contribute to item difficulty?

Rationale for examining linguistic and discousal features as predictors of item difficulty

RQ3 deals with the most concrete level of Chapelle et al. (2008) explanation planes that involves the examination of the surface features of tasks and how these features predict item difficulty. This is done by the quantification of linguistic and discourse features of the texts, items, and item-text variables of the tests and the statistical modelling of these features as predictors of the test item difficulty. In search of surface features – based evidence to support the explanation inference, that is the interpretation and use of the L-VSTEP reading test scores should be related to the theoretical construct of general reading proficiency, it is hypothesized that two sub-claims should be made: first, the linguistic and discourse constituents of the L-VSTEP reading test tasks should be congruent with the difficulty of the test items; and second, any linguistic and discourse constituents of the test tasks that are irrelevant to the construct of the test should contribute minimally to item difficulty.

Regarding the former, some criticism has been made of the use of passage-based reading comprehension tests, in that the validity of the interpretation and use of test scores can be seriously compromised if test-takers arrive at correct answers without even attending to the reading texts (Farr et al., 1990; Freedle & Kostin, 1999). In other words, comprehension of the reading texts – the actual construct of the test to be measured – is contaminated or disguised by other irrelevant factors such as prior knowledge or experience. It should be noted that prior knowledge or experience constitutes important dimensions in the top-down approach to reading comprehension, but the explanation of reading test scores based solely on these components clearly undermines the interpretation and use of the test scores. In order to provide evidence in alleviation of the above criticism, Freedle and Kostin (1999) argued that at least some correlation must be established between linguistic and discourse features of the texts and item difficulty.

Concerning the latter, in his seminal work on language test validity, Messick (1989) claimed that there were two major threats to the interpretation and use of test scores: construct irrelevant variance and construct underrepresentation. Construct irrelevant variance refers to the

variance in test scores accounted for by factors irrelevant to the constructs that would otherwise be intended to be measured by the test. On the other hand, construct underrepresentation indicates that tests with a mere narrow focus on small or unimportant facets underrepresent the constructs being measured. Construct irrelevant variance may derive from various sources among which test method, particularly multiple choice tests, has been given due consideration. Critics of multiple-choice questions claimed that students can perform at above-chance levels when they guess the answers without passages and that item difficulty is primarily governed by item information while text information plays less important roles (Freedle & Kostin, 1993). Gorin and Embretson (2006) argued that reading comprehension should be based on the difficulty and complexity of the texts rather than on the difficulty of the questions themselves. If the difficulty emanated from the questions contaminates reading comprehension by overriding the difficulty and complexity of the texts, then score-based inferences cannot be considered substantially valid. In addition, Gorin and Embretson (2006) further noticed that the test items in their study failed to assess readers' comprehension at text levels while the test's construct definition detailed a broad range of comprehension and reasoning processes, which was suggestive of construct underrepresentation.

In order to examine if test method (e.g., multiple choice) constitutes a construct irrelevant and construct underrepresentation factor to the L-VSTEP reading test, it is essential to explore the extent to which the texts have been comprehended and the extent to which this comprehension is substantiated by the text and item properties. If item difficulty correlates more strongly with item features than text features, then the test shows evidence of serious construct irrelevance. On the other hand, if text features account for more significant variance in item difficulty than do item features, the validity of the interpretation and use of the test scores can be supported.

Freedle and Kostin (1999) suggested that the validity of passage-based multiple-choice item comprehension tests should be supported by three different levels of evidence. At the lowest level, there should be some correlational evidence between text variables and reading test item difficulty. This is to dispel the most extreme criticism aimed at the validity of reading comprehension tests that test-takers pay no attention to the texts per se. At a higher level, in case of both text features and item features correlating significantly with item difficulty, there should be evidence demonstrating that text variables play more significant roles in explaining item difficulty. This can be done by using standard regression analysis in which all text and item variables enter the model simultaneously; then the relative contribution of each variable to item difficulty can be weighted

according to their associated Beta values. At the most rigorous level of evidence, it is essential to indicate that text variables remain a significant source of variance in the item difficulty even after partialling out the effects of item variables. Hierarchical regression analysis, in which each block of variables enters the model in a hierarchical manner, can be used to examine this hypothesis.

All three levels of evidence discussed above are examined in the current study. As such, data analysis for the study proceeds in three steps: the coding of text, item, and item-text variables, the estimation of item difficulty based on the Rasch model, and the statistical modelling of item difficulty and text, item, and item-text variables based on correlation and regression analyses. These are demonstrated in Figure 4.4.

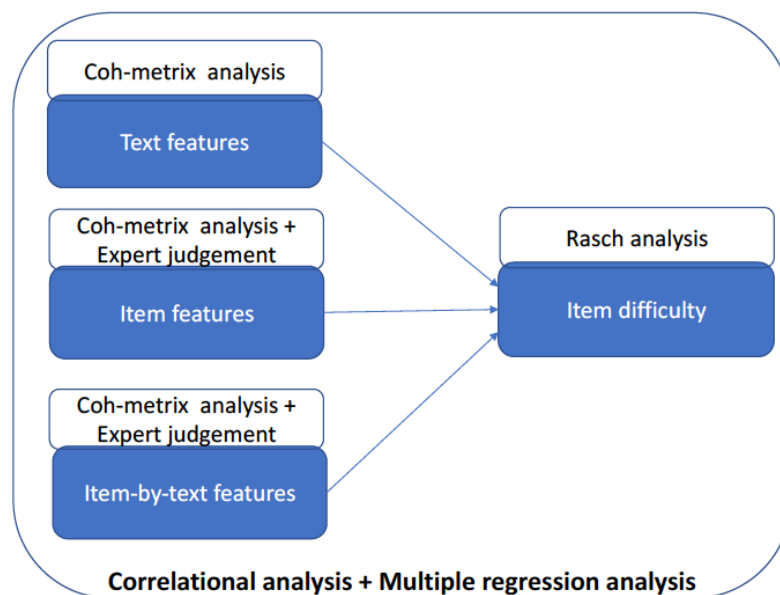


Figure 4. 4. Data analysis procedure for RQ₃

Participants

This phase of the research project involves participants responding to three different but equivalent forms of the L-VSTEP reading test, namely tests B, C, and D. Rather than each student taking three test forms repeatedly, they were assigned to three different groups, each taking one version of the test. In addition, to capitalize on the full information available from the project's data set and to enhance the robustness of the subsequent multiple regression analysis, data collected from form A of the test in the previous phase were also included in the analysis of this phase. To this end, the total number of students (N = 544) who completed form A of the test were assigned to four different groups, namely A, B, C, and D based on their scores obtained from form A. The assignment was conducted such that the distribution of high-, middle- and low-level

students and students with different academic disciplines was balanced across four groups. Students' levels of performance on form A of the test were determined on the basis of the cut-off scores adopted by the test administrators in assigning test-takers to different proficiency levels. Details are presented in table 4.5.

Table 4. 5. Distribution of the study participants.

Sample distribution by test score (form A), academic discipline, and gender (N = 544)				
	Group A (N = 136)	Group B (N = 136)	Group C (N = 136)	Group D (N = 136)
Test scores				
11 – 18 (B1/3)	35	35	35	35
19 – 31 (B2/4)	92	92	92	93
32 – 40 (C1/5)	9	9	9	8
Academic disciplines				
Pedagogy	36	36	36	35
Translation	53	53	53	53
Non-English	47	47	47	48
Gender				
Males	16	16	15	15
Females	120	120	121	121

Sample distribution by test score (forms A, B, C, and D), academic discipline, and gender				
	Group A (test A) (N = 136)	Group B (test B) (N = 123)	Group C (test C) (N = 84)	Group D (test D) (N = 104)
Test scores				
11 – 18 (B1/3)	35	44	33	33
19 – 31 (B2/4)	92	79	47	71
32 – 40 (C1/5)	9		4	
Academic disciplines				
Pedagogy	36	34	21	26
Translation	53	44	32	41

Non-English	47	45	31	37
Gender				
Males	16	11	7	9
Females	120	112	77	95

The majority of students who took form A of the test achieved scores at B2 Level (N = 369), followed by those at B1 Level (N = 140). Only a small number of students achieved scores at C1 Level (N = 35). Students in groups B, C, and D were respectively given test B, C, and D. The total number of students who turned up on the scheduled test date was 311, including 123 students in group B who took test B, 84 in group C who took test C, and 104 in group D who took test D. As can be seen in Table 4.5, a large proportion of students who took forms B, C, and D fell in the B1 (N = 110) and B2 (N = 197) Level score range, with only a few of them were in the C1 Level (N = 4). Given the inclusion of group A students in the final analysis, the total number of students in this phase was 447 (311 students who took test B, C, and D and 136 students who were assigned to group A but did not take any other tests).

Coding of the text, item and item-by-text variables

A major methodological component involved in addressing the research question was to code the linguistic and discourse features of the texts and items. To this end, the coding scheme supported by automated textual analysis tools and expert judgment were employed. Previous studies took the cognitive processing model for paragraph comprehension by Embretson & Wetzel (1987) as the primary approach for coding data. These coding schemes were based primarily on the manual count of word frequency and proposition density of texts or expert judgment of the degree of information overlap between the texts and test items, or among the stem and distractors of an item (Carr, 2006; Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gorin & Embretson, 2006) while numerous other linguistic and discourse features relevant to the reading texts such as text cohesion, syntactic complexity, text readability or vocabulary levels went unattended.

The development of computational linguistics, informed by corpus-based studies, has given rise to a number of computer software programs which help render the coding process less labour-intensive and more accurate. One such software is Coh-matrix. Coh-Matrix is a free online platform that generates indices of various linguistic and discourse features of texts including cohesion, vocabulary, syntactic complexity, and text readability; and therefore, provides a more

representative account of text and item features for the current study. In addition to automated textual analysis, expert judgment was also used for the coding of variables that went beyond the coverage of Coh-metrix. The identification and classification of text, item, and item-text variables for coding was gleaned from relevant theory and research reviewed in chapter II. Notable features that emerged from previous studies included those detailed in table 4.6.

Table 4. 6. Summary of text, item and item-text variables

	Variables	Methods of coding	Previous studies
Texts	Text length	Coh-metrix	(Gorin & Embretson, 2006), (Rupp, Garcia, & Jamieson, 2001)
	Syntactic complexity	Coh-metrix	(Ozuru et al., 2008), (Rupp et al., 2001), (Crossley, Greenfield, & McNamara, 2008), (Graesser et al., 2004)
	Lexical features	Coh-metrix	(Read, 2005), (Crossley et al., 2008), (Ozuru et al., 2008)
	Cohesion	Coh-metrix	(Graesser et al., 2004), (Ozuru et al., 2008), (Crossley et al., 2008),
	Text concreteness/abstractness	Coh-metrix	(Graesser, McNamara, & Kulikowich, 2011; Graesser et al., 2004)
	Text readability	Coh-metrix	(Graesser et al., 2011; Graesser et al., 2004; Green et al., 2010)
	Items	Item length	Coh-metrix
	Item lexical features	Coh-metrix	(Barkaoui, 2015)
	Degree of lexical overlap between correct answer and distractors	Coh-metrix	(Freedle & Kostin, 1993)

Item-by-text	Number of plausible distractors	Human judgment	(Ozuru et al., 2008; Rupp et al., 2001)
	Level of abstractness of question	Coh-metrix	(Ozuru et al., 2008; Rupp et al., 2001)

Text variables

Text variables that were found to be significant contributors to item difficulty in studies reviewed in chapter II were selected for the current study. These included text length, syntactic complexity, lexical features, cohesion, text concreteness, and text readability.

Text length

Text length indicates the total number of words in a text. Longer texts may impose heavier cognitive load on working memory, making it more difficult for cognitive processing than do shorter texts (Gorin & Embretson, 2006; Rupp et al., 2001).

Syntactic complexity

Syntactic complexity refers to the range and sophistication of forms that occur in texts (Ortega, 2003). Several indices of syntactic complexity generated by Coh-metrix were examined: Sentence length is measured by the average number of words per sentence in a text. Left embeddedness is calculated as the number of words preceding the main verbs. Noun phrase density which refers to the mean number of modifiers per noun phrase is also calculated. It is hypothesized that shorter sentences, fewer words before the main verbs and fewer modifiers in a noun phrase make syntactic parsing easier (Crossley, Allen, & McNamara, 2012; Just & Carpenter, 1992; Mcnamara, Graesser, McCarthy, & Cai, 2014).

Lexical features

The difficulty of texts is influenced by various lexical features among which lexical diversity, word frequency, and word familiarity were examined in the study. Lexical diversity, usually referred to as type-token ratio (TTR), is the ratio of unique words (Type) to the total number of words in a text (Token). The higher the ratio, the larger the number of different words introduced in a text, and the more difficult it is to process the text, thus contributing to item difficulty (Crossley et al., 2012). Since type-token ratio is mediated by text length (Koizumi, 2012; Malvern & Richards, 2002; McCarthy & Jarvis, 2010) the current study reported a similar index

generated by Coh-metrix – The Measure of Textual Lexical Diversity (MTLD) – which is designed to counteracts the confound effect of text length.

Word frequency concerns the frequency with which a word occurs in a text. Frequent words aid the decoding and processing of texts, and are linked to “richer bodies of world knowledge” (Beck, McKeown, & Kucan, 2002; Perfetti, 2007). On the other hand, rare words render text decoding and processing more difficult, hence representing a major limiting factor in text comprehension (Graesser et al., 2011). The frequency count in Coh-metrix is based primarily on the database from the Centre for Lexical Information (CELEX) which includes frequencies derived from the 1991 version of the COBUILD corpus (Crossley et al., 2012; Crossley et al., 2008).

Word familiarity is the degree of familiarity with which an adult conceives of a word. The more familiar the words in a text are, the more quickly the text is processed (Mcnamara et al., 2014). The rating of word familiarity in Coh-metrix is based on the Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981) which consists of ratings for 150.837 unique words along several psychological dimensions (Graesser et al., 2004; Mcnamara et al., 2014).

Cohesion

Cohesion refers to the “explicit features, words, phrases, or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas and in connecting ideas to higher level global units” (Graesser et al., 2004). Following the works of Barkaoui (2015), Green et al. (2010) and Graesser et al. (2004), three forms of cohesion were examined in the study, namely referential cohesion, conceptual cohesion and connective density.

Referential cohesion refers to the degree of coreference between words in a text. Coh-metrix yields several indices of referential cohesion with varying degree of overlap such as overlap at local (adjacent sentences) and global (across sentences) levels, and overlap with respect to different types of coreference (noun overlap, argument overlap, stem overlap and content word overlap). Content word overlap was found to contribute significantly to item difficulty in Crossley et al. (2008) study, while stem overlap, noun overlap and argument overlap constituted major prediction of material adaptation in Crossley, Louwerse, McCarthy, and McNamara (2007) and Crossley et al. (2008) studies. These forms of coreference at global level (across all sentences in a text) were reported in the current study.

Conceptual cohesion regards the levels of semantic and conceptual similarity between sentences or paragraphs in a text. Text cohesion is expected to increase in accordance with the

level of semantic and conceptual similarity between text constituents (Crossley et al., 2008; Graesser et al., 2004). Coh-matrix computes measures of Latent Semantic Analysis (LSA) as indicators of conceptual cohesion on a scale of 0 to 1, with higher scores indicating higher semantic similarity between sentences or paragraphs in a text. The current study reported two measures of LSA: LSA sentence adjacent – semantic similarity between two adjacent sentences and LSA sentence all – semantic similarity between all sentences across the text.

Connectives density provides important links between ideas and clauses in a text and evinces clues about text organization (Crossley et al., 2007; Mcnamara et al., 2014). Coh-matrix provides incidence scores (occurrence per 1000 words) for different classes of cohesion identified by Halliday & Hasan (1976) and Louwse (2001) such as causal (because, so), logical (and, or), adversative/contrastive (although, whereas), temporal (first, until), and additive (and, moreover).

Text concreteness / text abstractness

The level of abstractness/concreteness of texts is believed and has been found to contribute significantly to item difficulty (Freedle & Kostin, 1993). Coh-matrix incorporates two lexical databases – the MRC psycholinguistic database (Coltheart, 1981) and the Wordnet database (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) – to generate measures indicative of text abstractness/concreteness. One such measure is word concreteness computed as z-score for each text. The higher the z-score value, the higher the portion of meaningful and concrete words, and the lower the processing demand (Barkaoui, 2015). This is also the text abstractness/concreteness index reported in the current study.

Text readability

Text readability provides initial and shallow measures of text difficulty based primarily on word length (the number of either letters or syllables per word) and sentence length (the average number of words per sentence in a text). Three common readability formulas are computed in Coh-matrix: the Flesch Reading Ease, the Flesch Kincaid Grade Level and the Coh-matrix L2 readability. The Flesch Reading Ease scores range from 0 to 100 with higher scores signifying more difficult texts. On the other hand, the Flesch Kincaid Grade Levels formula produces scores tailored to the US grade-school level, with higher scores indicating more difficult texts. The third readability formula – the Coh-matrix L2 readability – considers not only text complexity at word and sentence levels, but also the cohesion between sentences in the text. Crossley et al. (2008) and Crossley et al. (2012) found that the Coh-matrix L2 readability represented a more significant

predictor of L2 readers' performance on academic texts, and of text level classification than the other two readability formulas. However, only the Flesch Reading Ease scores were reported in the current study because the Flesch – Kincaid Grade Level is more subject to the US grade system while the Coh-metrix L2 readability formula includes variables that are already computed in other Coh-metrix indices.

Item variables

Item length and item vocabulary

Item length refers to the total number of words in stems and all options. Two measures of item vocabulary were examined: word familiarity and word frequency. Just as text length and text vocabulary impose higher processing demand on text comprehension, it is expected that longer stems and options with more unfamiliar words and more rare words make item processing more demanding, thereby adding another layer of cognitive load onto readers' answering comprehension questions.

Lexical overlap between correct answers and distractors

This variable is computed as the proportion of content word overlap between the correct answer and the three distractors for a given item. Items with higher content word overlap are more difficult to process than items with lower content word overlap (Freedle & Kostin, 1993).

Item-text variables

Number of plausible distractors

Rupp et al. (2001) proposed a coding scheme of plausible distractors for one-stem-four-option multiple choice item tests. As such, the number of plausible distractors (out of three) is determined by the propositional overlap between the information in the distractors and the information in the texts. A distractor is considered plausible if it contains ideas that are either directly stated in the text or indirectly inferred from the information given in the text. Items with more plausible distractors make it more difficult for the item confirmation and falsification process.

Lexical overlap between the text and the correct option

This variable concerns the overlap between words in the correct answer and the text. Lexical overlap between the text and the correct answer for a given item is a major predictor of item difficulty in Freedle and Kostin (1993, 1999). Among different indices of word overlap computed by Coh-metrix, the content word overlap index was reported in the current study.

All item, text, and item-text variables identified above are treated as continuous variables measured on an interval/ratio scale. For the four item-level variable that requires subjective judgment on the number of plausible distractors, two independent coders were employed to conduct independent coding across all 40 items of each test version. Face-to-face discussion was then be conducted to ensure that coding agreement reaches satisfactory levels and all disagreements were settled before the statistical analysis process.

Rasch analysis

In order to examine the relative contribution of text, item, and item-by-text variables to the item difficulty of the tests, it is essential to statistically produce appropriate measures of item difficulty to serve as the criterion variable in the subsequent regression analyses. Most previous studies on item difficulty modelling used classical test theory (CTT) procedures to estimate item difficulty. However, one disadvantage of CTT is that it is subject to the specific sample on which data are modelled (Barkaoui, 2015; Knoch & McNamara, 2015). In other words, CTT may yield different measures of item difficulty if it is applied to a different group of learners in a different context. For this reason, this study employed the Rasch model as a subset of the item response theory (IRT) models to generate the item difficulty index. Rasch analysis takes sample dependency into account, and therefore produces results that are generalizable (Knoch & McNamara, 2015).

Rasch analysis is a probability-based measure of success of a particular test-taker on a given item. This probability of success is examined by using an equal interval scale measured in logits (or log odd units). This equal interval scale assumes equal values for equal distances at anywhere on that scale, thus allowing for direct comparison of item difficulties and person abilities (Green, 2013). The likelihood of a test-taker answering an item correctly depends on how well the measures of that test-taker's ability and the given item difficulty match up against each other on a common equal interval scale (Knoch & McNamara, 2015). There are four different models of Rasch with respect to different types of measurement instruments and different data types, namely the simple Rasch model, the rating scale Rasch model, the partial credit Rasch model, and the many-facet Rasch model (Bond & Fox, 2015). Since the L-VSTEP test is composed of all dichotomously-scored items, the simple Rasch model was employed.

The simple Rasch model generates indices of item difficulty, person ability and a Wright map that plots both item difficulty and person ability measures onto a common scale. This pictorial representation of the relative standing of each test-taker in relation to the difficulty of the test items

is one of the most interesting pieces of information provided by the Rasch model as it instantly offers a visual inspection of how easy or difficult the test is for a given group of learners. In addition, indices of fit statistics, reliability, and separation are also provided, each of which is reported as essential information accompanying the item difficulty index, and is briefly described below.

Fit statistics refers to the discrepancy between the person-item matrix predicted by the Rasch model and the person-item matrix representing the actual observed scores from the test (Bond & Fox, 2015). This statistics is usually expressed in terms of two chi-square ratios: infit mean square and outfit mean square (MNSQ). Infit mean square is an information-weighted index exhibiting unexpected patterns of scores that are close to the mean item difficulty. On the other hand, outfit mean square is sensitive to the detection of outliers that deviate from the expected score patterns (Aryadoust, 2013; Bond & Fox, 2015; Green, 2013). Misfit items are those considered to be ill-represented by the test instrument, and can be detected by how much they depart from a certain range of fit statistics. This study follows Bond and Fox (2015) in interpreting any items that are outside the range of 0.6 and 1.4 as misfitting items. Fit statistics can also be expressed in standardized forms as ZSTD infit and outfit with the z scores outside the ± 1.96 range indicating misfit items.

Rasch provides indices of both item reliability and person reliability. Item reliability concerns the levels of confidence with which the Rasch model can yield the same measure of item difficulty given another group of test-takers of the same size and ability and under the same conditions. Similarly, person reliability refers to the consistency with which the relative standing of a person on the logit scale can be replicated if he/she is given another test of the same difficulty level. The higher the reliability values, the more confident one can have in the precision of item difficulty and person ability measures (Green, 2013). Rasch reliability can also be expressed in terms of item and person separation indices. Item and person separation indices can be used to ascertain whether there are enough items to reliably distinguish performers of different levels; and whether the sample size is large enough to reliably confirm item difficulty hierarchy (Linacre, 2014). Low person separation (< 2) with person reliability $< .8$ implies that the test does not discriminate well among high and low performers. On the other hand, low item separation (< 3) with item reliability $< .9$ indicates that the sample is not large enough to reproduce item difficulty hierarchy (Linacre, 2014).

Correlational analysis

Correlational analysis is a simple statistical technique used to explore if there is a linear relationship between two variables and how strong that relationship, if any, is. The probability value (p value) is usually set at .05 as the cutoff value for the determination of statistically significant correlation between two variables, while the *correlation coefficient* value (r) on a scale of -1 to 1 indicates the strength with which two variables correlate with each other. Guidelines for the interpretation of the strength of the *correlation coefficient* are as followed:

.70 < r < 1.00 : strong relationship

.40 < r < .70 : medium relationship

.10 < r < .40 : weak relationship (Cohen, 1988)

Correlation coefficient can take on either negative or positive values. A negative value implies that two variable systematically move in the same directions so that increase in value of one variable is accompanied by increase in value of the other (Phakiti & Roever, 2018). On the other hand, a negative value indicates that two variables linearly move in opposite directions whereby increase in value of one variable is associated with decrease in value of the other. There are different types of correlation depending on the nature of the data set. Since all variables in the study are treated as continuous variables measured on interval or interval-like scales, the *Pearson Product Moment Correlation* (Pearson's r) was computed. Prior to computing Pearson's r , several statistical assumptions must be addressed such as the data set must be normally distributed and are spread cross a wide range or there are no extreme scores. These assumptions can be addressed through the inspection of descriptive statistics and the creation of a histogram (Phakiti, 2014). Spearman correlation was employed if the assumption of normal distribution was violated.

Correlational analysis was used to examine the degree to which text, item and item-by-text variables correlate with each other and with the item difficulty measure of the tests. This is the first step in the three-step approach to item difficulty modelling suggested by Freedle and Kostin (1999). The main purpose of this step is to provide evidence against the most extreme level of criticism regarding the validity of the test that test-takers pay little or no attention to the texts while answering the questions. Accordingly, it is expected that text and item-by-text variables should at least have significant correlation with item difficulty. In addition, this step also helps identify eligible variables - those having significant correlation with item difficulty and non-colinear relationship (variables correlating .90 or more) with other predictor variables - to serve as the

predictor variables for the subsequent regression analysis. Colinearity is of concern because if two variables are colinearly related ($r > .90$), it is impossible to compute unique estimates of the regression coefficient for each of them (Field, 2009).

Multiple regression analysis

Multiple regression analysis is a correlation-based statistical procedure used to model the relationship between a single criterion variable (or dependent variable) and multiple predictor variables (Independent variables) via the creation and estimation of a prediction equation that best accounts for that relationship (Ho, 2014).

Following the three-step approach to item difficulty modelling suggested by Freedle and Kostin (1999), both standard regression and hierarchical regression are employed in the current study. Standard regression is performed on promising predictor variables extracted from the first step to determine if text variables play a more significant role than item variables in explaining item difficulty. It is expected that one or more text and item-text variables should account for more variance in item difficulty than do item variables. After that, hierarchical regression is computed to ascertain that text variables remain significant predictors of item difficulty even after the effects of item features have been partialled out.

Results were interpreted according to the essential information provided by the yielded values of unstandardized regression coefficient (B), standardized regression coefficient (β), multiple correlation coefficient (R), R square (R^2), adjusted R square ($adj R^2$), and the statistical significance F -change test. Unstandardized regression coefficients B carry the weight of each predictor variable in the regression model. It shows the predicted amount of change in the value of the criterion variable given a one-unit change in the value of the associated predictors while holding constant the value of other predictors. The standardized regression coefficient β is interpreted in the same manner as the unstandardized regression coefficient but is obtained when all variables are standardized so that each has a standard deviation of 1 and a mean of zero (O'Rourke et al., 2005). In terms of weighting the relative importance of predictor variables, it is more useful to report the β value than the B value because the former can enable comparison of relative importance across predictor variables (O'Rourke et al., 2005; Phakiti & Roever, 2018). The R value denotes the multiple correlation coefficient between the criterion variable and the predictor variables combined, and is interpreted in the same way as the Pearson Product Moment Correlation. The squared value of R (or R^2) illustrates the amount of variance in the criterion

variable collectively constituted by the predictor variables. The higher the value the more variance explained. Adjusted R square ($\text{Adj } R^2$) takes into account the sample size and yields values germane to the interpretation of the model as it is generalized to the population. Both R square and Adjusted R^2 were reported in the current study. The significance F -change test expresses how well different regression models with nested structure explain variance in the criterion variables. The F -change significance value is of particular importance in comparing different models with nested structure in a hierarchical regression analysis.

Results of multiple regression analysis can be seriously affected by various factors among which, effects of influential cases and multicollinearity are the most oft-noted. The former refers to extreme cases that exert undue influence on the regression model as a whole, whereas the latter concerns the excessive correlation among predictor variables that renders the unique estimates of regression coefficients impossible (Field, 2009). In the current study, these factors were respectively managed by the inspection of the values of Mahalanobis distance and the variance inflation factors (VIF) (Field, 2009).

4.3.4. Research question 4

To what extent do students' scores on the L-VSTEP reading test predict their reading performance in the relevant academic programs?

In order to generalize students' observed scores beyond the test itself and into the real-world situation, the comparability between the test domain and the target language use domain should be established to support the extrapolation inference. The research question addressed in this phase provides one such evidence: the extent to which students' scores on the test predict their performance in the relevant target language use domains. To the extent that the predictive relationship between students' test scores and their expected scores in the target language use domain is not supported by empirical evidence, the extrapolation inference is weakened. The question remains as to what measure can provide a comprehensive assessment of students' performance in the target language use domain. Previous studies employed different potential measures, such as well-established standardized proficiency tests, grade point average, teacher's evaluation, and students' self-assessment of their own ability (Li, 2015b). The current study sets out to use student self-assessment as a measure of students' performance in the target language use domain for several reasons. First, since constant self-reflection is a typical characteristic of human beings, self-assessment enables students to provide a more comprehensive and accurate

evaluation of their performance than any other external measures, making them a proactive assessor of their own learning rather than the receiver of assessment (Fan & Yan, 2017; Powers & Powers, 2015). Second, self-assessment has the potential to allow students to reflect on multiple indicators of their own performance in a diverse range of contexts, thus providing more accurate assessment of their own learning as compared with the limited scope of exams (Upshur, 1975). Last but not least, self-assessment is easier to design, administration, and scoring as compared with teacher’s evaluation or standardized tests (Brown, Dewey, & Cox, 2014). Figure 4.5 provides a graphical representation of the data analysis for this stage.

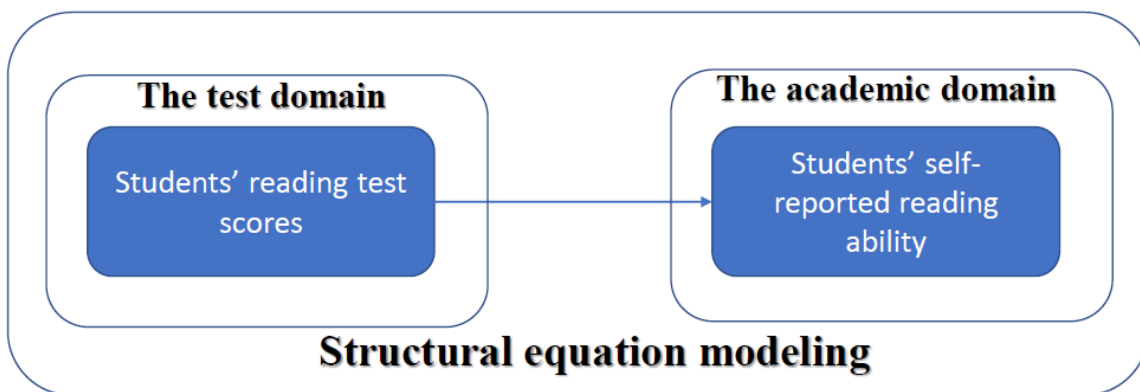


Figure 4. 5. The data analysis procedure for RQ4

Participants

The same 544 students who took the test version A (see section 4.3.2) were invited to answer the questionnaire one week before they took the test. Out of the 544 students who were invited, 344 students completed and returned the questionnaire. Results of a t-test suggested that there was no statistically significant difference between the sample of 344 students and the sample of 544 students in terms of their scores on the test version A. With regards to participants’ disciplines, 133 students (38.7%) were majoring in English teacher education, 88 (25.6%) were majoring in English translation/interpretation, and 123 (35.8%) were non-English major students with a majority in the social science areas. They were aged between 20 and 22 (M = 21.13, SD = 0.79); 9.89 % (N = 34) were male and 90.11% (N = 310) were female. The male-female disproportion as reported above could be explained by the fact that the majority of students who gave consents to participate in the study were primarily from social science disciplines where there was an overrepresentation of female students.

Development of the self-assessment questionnaire

The first phase of this study sought to develop and validate a reading proficiency self-assessment scale for the study participants. In the second phase, a structural equation model that captures the predictive relationship between students' test scores and their self-reported English reading proficiency was proposed and tested.

The self-assessment scale items were developed based on a) a thorough review of the extant literature on L2 reading comprehension in chapter II, b) the Vietnamese version of the Common European Framework of Reference for languages (CEFR-VN), c) the guidelines for the development of the L-VSTEP reading test (Ministry of Education and Training of Vietnam, 2015a; Nguyen, 2018), d) the L2 reading comprehension curricula adopted at the institution where the data collection took place, and e) DeVellis (2016) suggestions for best practices in scale construction and validation.

As discussed above, the CEFR-VN includes language proficiency descriptors of six levels of reference compatible with those in the CEFR of the Association of Language Testers in Europe and the Council of Europe (Nguyen & Hamid, 2015). For the purpose of this study, however, only reading proficiency descriptors at level 3/B1 to level 5/C1 were considered. Reading proficiency of the Vietnamese EFL learners, as described in the CEFR-VN at level 3/B1 to level 5/C1, is generally characterized as the ability to locate and understand explicit ideas and arguments, to infer implicit meaning from the texts, and to summarize textual information (Ministry of Education and Training of Vietnam (MOET), 2014).

Following Oscarson (1997) and Bandura (2006) suggestions for writing self-assessment items, the guidelines for the development of the L-VSTEP reading test and the L2 reading curriculum employed at the corresponding institution were also consulted to operationalize the constructs of reading proficiency as defined in the CEFR-VN, and to develop a more contextualized and criterion-referenced self-assessment instrument relevant to the learning experience of the targeted participants. For example, one of the reading proficiency descriptors at level 5 in the Framework was the ability to understand subtle details such as the implied meaning of a detail/argument in the text or the attitude of the text's author. This proficiency description was further specified in the test development guidelines as the ability to understand the tone and attitude of the author in a detail or argument in the text, the ability to decipher the message that the author wants to convey via a detail in the text, or the ability to understand the logical inferences/arguments

of the text's author. Similarly, these abilities were operationalized in the L2 reading curriculum as specific subskills embedded in each learning unit that students need to employ in the interpretation and comprehension of the various reading passages and lessons.

This process engendered an initial item pool of 32 items which was then further refined to rule out redundant, ambiguous and double-barreled items (DeVellis, 2016). For example, the item "I can synthesize information across the text to make inferences" was deleted since it involved two different subskills, integrating information and making inferences, which made it difficult for the students to discern. A 27-item scale was finally created and ready for the content scrutiny phase. Vietnamese was the language of choice for the questionnaire items as the CEFR-VN and the guidelines for the L-VSTEP test were both written in Vietnamese. A six-point Likert scale was adopted, wherein participants express their degree of agreement with the statements via their endorsement of the six categories (from strongly disagree to strongly agree). Two language testing experts who were trained as the L-VSTEP test item writers under the 2020 project and who were responsible for test design at the institutional level participated in this phase. Each of them independently and then collectively scrutinized each of the 27 items in terms of their content, wording, clarity and completeness. None of the 27 items was dropped during this phase but several items were suggested to be reworded to ensure the intelligibility, content representativeness, and clarity. Nine students were then invited to respond to the 27-item scale. In light of their comments, some further modifications were made. For example, the words "*từ qui chiếu*" (*referent words*), "*hàm ngôn*" (*implicit meaning*) and "*hiển ngôn*" (*explicit meaning*), which have Chinese origin and were confusing for students, were clarified by relevant examples and explanations in a separate section of the questionnaire. The refined questionnaire was then piloted with an intact class of 43 students to check for its internal consistency. A Cronbach's alpha of 0.934 was yielded, rendering the questionnaire appropriate for the main data collection stage.

The data analysis in the study proceeds in three steps: Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for the validation of the self-assessment questionnaire and structural equation modeling to address the research question.

The whole data set was initially subjected to an exploratory factor analysis, using IBM SPSS 22 software, to unearth the underlying factor structure of the self-assessment scale (Field, 2009; Hair et al., 2014; Loewen & Gonulal, 2015). Before running EFA, the assumptions of sampling adequacy and variable correlation were checked, using the Kaiser – Meyer – Olkin

measure of sampling adequacy (KMO) (Kaiser, 1970) and the Barlett's Test of Sphericity respectively. A KMO value higher than 0.5 indicates that the sample size is large enough to produce stable factor solution, while a significant Barlett's Test of Sphericity is suggestive of sufficient intercorrelation among variables (neither too high, nor too low) (Field, 2009). Principle Axis Factoring was set as the factor extraction method as it does not assume multivariate normal distribution of data (Fabrigar & Wegener, 2011). Promax rotation method was chosen with the assumption that the underlying factors were correlated to explain participants' response to the self-assessment scale (Field, 2009). To determine the optimum number of factors to be retained after extraction, multiple criteria including the Kaiser's criterion (eigenvalues higher than 1), the Scree plot, the cumulative percentage of variance, and the interpretability of the extracted factors were used in conjunction. Factor loadings of 0.3 or higher were set as indicators of the substantive importance of variables (scale items) to a given factor (Field, 2009). In addition, cross-loading variables – variables with high loadings on more than one factor were also given special attention.

The procedure for conducting Confirmatory Factor Analysis and Structural Equation Modeling is similar to that discussed in section 4.3.2. The same five stages of model specification, model identification, model estimation, model evaluation, and model modification as well as the global model fit indices are adopted for the evaluation of the two analyses. Of most importance in the Structural Equation Modeling, however, is the regression coefficient that indicates the structural relationship between students' test data and self-assessment data. A significant value of the regression coefficient suggests that the students' test scores significantly predict their self-assessment scores while the magnitude of the predictive relationship is indicated by the Beta value.

4.3.5. Research question 5

To what extent are reading tasks and skills assessed in the L-VSTEP reading test aligned with reading tasks and skills required in the relevant academic programs?

The question addressed in this study provides additional evidence for the extrapolation inference which makes a claim about the relationship between students' observed performance on the test and their expected performance in the target language use domain. The study takes a task-centered perspective to examining the comparability between the reading tasks and skills assessed in the L-VSTEP reading test and those considered important in the relevant academic programs. If the reading tasks and skills sampled in the test provide a good coverage of the reading tasks and skills required in the academic programs, the extrapolation inference can be considered plausible.

Otherwise, the extrapolation inference would be weakened because the test domain underrepresents the target language use domain. To this end, semi-structured interviews were conducted with both lecturers and graduate students who have extensive experience with both domains, the L-VSTEP reading test and the academic programs at the relevant institution. Figure 4.6. is a graphical summary of the research procedure for RQ₅.

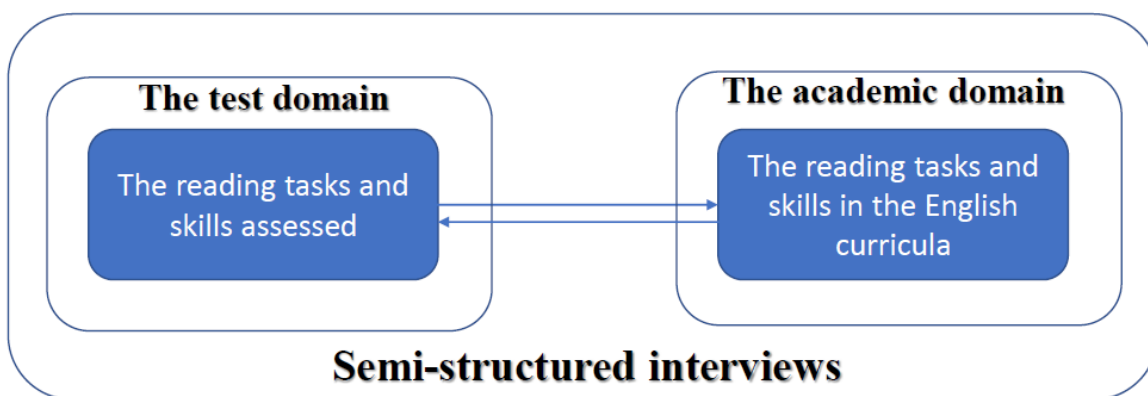


Figure 4. 6. The data analysis procedure for RQ₅

Participants

In order to gain insights into key stakeholders' experience with the language demands of the relevant academic settings in terms English reading tasks and activities as well as their perception about how comparable these tasks and activities to those found in the test, three lecturers of English and three newly-graduated students at UA were invited to participate in a semi-structured interview. In what follows a brief description of each interview participant including their professional and academic background as well as their experience with the English curricula and the L-VSTEP reading test at the institution is presented to warrant their eligibility for the study. To adhere to the ethical guidelines, pseudonyms are used for all the informants.

Lecturer A (LA) is an English lecturer who holds a Master's degree in Linguistics with more than seventeen years of teaching experience at the institution. LA has been primarily involved in the development and offering of English courses to both English major and non-English major students in a variety of subjects, including English morphology, English pronunciation training, English reading, listening, translation and interpretation for English majors as well as general English communication (English 1, 2, 3 courses) for Non-English major students. At the time of interview, LA was in charge of teaching English translation to four English for translation and interpretation classes, and English pronunciation training and English reading

skills to three English pedagogical classes. LA was also involved in the delivery of the L-VSTEP test preparation classes to students across different academic disciplines at the university English centre.

Lecturer B (LB) is an English lecturer with a Master's degree in Linguistics and more than 24 years of teaching experience at the institution. LB is among a few lecturers who are primarily involved in the design and teaching of English reading skills only to English major students, and so can be considered to have rich experience in teaching this particular macro skill at the institution. In addition, LB also had some experience delivering courses in British and American culture and literature for English major students. LB was also mandated by the institution to take part in training courses for the VSTEP test item writers and assessors offered by the Ministry of Education and Training and was responsible for organizing similar training courses at the institutional level. At the time of interview, LB was involved in the development of the L-VSTEP reading test and offering L-VSTEP test preparation classes for students across different disciplines at the university English centre.

Lecturer C (LC) is an English lecturer with a Master's degree in Linguistics and more than 26 years of teaching experience at the institution. LC has been involved in the design, redesign, and renovation of English for specific purposes curricula for many years at the institution, and hence had insightful knowledge and rich experience with teaching non-English major students. LC has also been teaching English translation and interpretation to English major students and supervised students' BA theses in these areas. At the time of interview, LC was teaching English language skills and grammar to English major students as well as English for finance and banking, English for physical education, English for engineering and technology, and English for geography to non-English major students in the respective faculties.

Student A (SA) is a fresh graduate from the institution majoring in English teaching. SA has completed all academic modules, took the L-VSTEP test and achieved scores that satisfied the requirements in terms of English language proficiency for graduation. At the time of interview, SA had just received a job offer as a high school English teacher in SA's hometown.

Student B (SB) graduated from the same institution with a Bachelor's degree in mathematics teaching. SB has satisfied all requirements including sufficient L-VSTEP scores before graduation. At the time of interview, SB was a maths teacher at a high school in the same province.

Student C (SC) graduated from the same institution with a Bachelor’s degree in English language (English for translation and interpretation), meeting all required criteria for graduation including satisfactory L-VSTEP scores. At the time of interview, SC was a freelance teacher and a registered translator for a government agency.

Data collection and data analysis

Semi-structured interviews were conducted with each individual teacher and student at a time and place convenient for them. A set of interview questions were sent to the participants before the interview scheduled dates and they were encouraged to review the teaching and learning materials that they have been using so as to provide informed answers to the interview questions. The interview questions revolve around three main aspects of relevance to the comparability between the reading tasks, activities, and requirements in the academic domains and those in the test. These includes the amount and type of reading required of students in the undergraduate programs, the reading skills considered important to perform adequately in the undergraduate programs, and the alignment between the academic domains and the L-VSTEP reading tests in terms of reading tasks, skills, and difficulty levels. These interview questions served as the core questions that covered the focus of the research question while spontaneous follow-up interview questions were used by the researcher during the interviews to probe further elaborations and clarifications from the informants as new developments of the stories emerged, thereby capitalizing on the benefits that a semi-structured interview approach offers (Dornyei, 2007; Riazi, 2016). Table 4.7 presents the questions for the semi-structured interviews.

Table 4. 7. Interview questions

Amount and types of reading required	Reading skills required and reading tasks commonly encountered	Comparability between the academic domains and reading tests.
1. How important is English reading to undergraduate students?	1. What reading skills are important for undergraduate students?	1. How comparable is the reading tasks required in the academic domains to those in the test?
2. How much reading are students required to do in their undergraduate programs?	2. What reading skills are students generally good at?	2. How comparable is the reading skills/abilities

<p>3. What types of reading are students required to do in their undergraduate programs?</p>	<p>3. What reading skills do students have difficulty with?</p> <p>4. What reading skills/abilities are students assumed to achieve by the end of their undergraduate programs?</p> <p>5. What reading tasks do students commonly encounter in the undergraduate programs?</p>	<p>required in the academic domains to the test?</p> <p>3. How comparable is the perceived readability of the reading texts in the academic domains to those in the test?</p>
--	--	---

Information related to the purposes of the study, data collection procedures, and ethical considerations were communicated to the participants before informed consents were obtained and schedules for each individual interview were agreed upon. Each interview session lasted for 30 to 45 minutes on average and was audio-recorded. Vietnamese was the main medium of communication, though some English was also allowed when participants found it more comfortable or hard to find Vietnamese equivalents of an English concept. Interview data were transcribed verbatim and relevant portions to be reported were translated into English by the researcher.

Interview data were thematically analysed in three phases. First, the researcher listened to the recordings and read the transcribed interviews several times to obtain a general sense of the interview data. Salient extracts from the participants' accounts that were relevant to the topics of interest were given special codes. These codes across different participants' accounts were later revisited and codes that were similar or closely related were regrouped to form clustered of categories that informed the understanding of the research problems. This iterative process went on until specific extracts from participants' accounts fitted into the broader categories. In the final phase, a researcher's colleague was invited to act as an independent coder who coded a portion of

the interview data using the categories identified in the previous phase. Any emerging disagreements were then discussed in a moderation session until final agreement was achieved.

This chapter provides an overview of the research method paradigm adopted to address the proposed research questions as well as a detailed elaboration on the procedure for data collection and data analysis to answer each of the research questions. In chapters V, VI, VII, VIII, and IX that follow, the findings derived from the research programs in response to each Research Question are respectively presented and relevant discussion sections are offered by drawing on the related theoretical and empirical literature.

CHAPTER V: EXPERT JUDGMENT AND STUDENTS' REPORTED READING PROCESSES

5.1. Introduction

This chapter reports the findings of the first research question: “What reading processes are assumed to correctly answer L-VSTEP reading test items? To what extent do these processes correspond with the reading processes actually engaged by test-takers while they take the test?”. Exploration of this alignment contributes to understanding the relationship between the construct the test is designed to measure and test-takers’ performance on the test. The specific aspect of test construct investigated in this section involves the comparison of the reading processes believed by the experts to be activated by the test items and the actual reading processes engaged by test-takers. The data collection and data analysis procedures were conducted in two separate steps: expert judgment on the reading processes that inform responses to the test items; and elicitation and analysis of the test-takers’ verbal reports. Findings of this research phase are presented in the following sections.

5.2. Findings from expert judgment

5.2.1. Findings from the pilot stage

During the piloting phase, the two experts were given a sample reading test comparable to the L-VSTEP test and the initial expert judgment form. Each of them read through the reading passages and answered the questions before consulting the answer key. They then worked collaboratively to discuss the answers and match the test items with the reading skills in the judgment form, following the below procedures:

- 1) Match a single item with a single reading skill primarily assessed by that item.
- 2) Identify potential involvement of other skills in response to that item.

The ultimate goal of this practice was to enable the experts to reach an agreement on the interpretation of the skill descriptions and to familiarize them with the judgment process. This practice resulted in some changes to the expert judgment form and the judgment procedure. First, as suggested by the two experts, the skill descriptions derived from the test development guidelines were too general, which made it difficult to interpret and apply in the judgment process. Therefore,

the skill descriptions were further refined exclusively for the purpose of the study, based on relevant skill definitions and descriptions in previous studies (Buck et al., 1997; Gao & Rogers, 2011; Kim, 2015). For example, the skill Understanding Cohesive Devices was further refined by providing more detailed descriptions and examples of the cohesive devices to facilitate experts' interpretations. Second, the wording of the initial translated version of the expert judgment form that might cause confusion among the experts was modified by using more precise terms and descriptions. For example, the skill "Understanding Specific Information" was replaced by "Understanding Explicit Information at the local level" to differentiate it from the skill "Integrating Textual Information" which also requires the understanding of details but at the global level; the skill "Understanding Rhetorical Information" was replaced by "Understanding Pragmatic Meaning"; and the skill "Identifying References" was replaced by the clearer term "Understanding Cohesive Devices". Third, the experts suggested that different items might test the same skill; yet depending on the level of difficulty of the item and the reading passage to which it belongs, the description of the skill might be different. Therefore, the descriptions of the skill were further refined to reflect the level of complexity with which it was used to answer a specific item. Details of the reading skills, their definitions and descriptions used in the main study are presented in Table 5.1.

Table 5. 1. Reading skills, definitions and descriptions

Skill	Definition	Description
Understanding explicit information at local level (UEI)	The ability to locate and understand explicit meaning at the sentence level.	<ul style="list-style-type: none"> - Understand specific details that are explicitly stated in the texts, using simple grammatical structures and vocabulary - Locate a specific detail in the text. - Identify and understand paraphrased information explicitly stated in the text.
Understanding cohesive devices (UCD)	The ability to understand the relationship between sentences or ideas using connective devices such as discourse markers, anaphoric and cataphoric references, substitutions, repetitions.	<ul style="list-style-type: none"> - Identify the antecedent of a pronoun. - Understand logical ideas in the text based on linking devices such as referent words, conjunctions, linking words, and repeated words.

Integrating textual information (ITI)	The ability to synthesize information from different parts of a paragraph or a text.	<ul style="list-style-type: none"> - Locate and synthesize information across a paragraph. - Locate and synthesize information across the text.
Summarizing textual information (STI)	The ability to understand main ideas and recognize supporting details at paragraph and discourse level.	<ul style="list-style-type: none"> - Understand the main idea of a paragraph. - Understand the main idea of a text - Identify and understand supporting details for an argument or the main ideas of a paragraph or a text.
Inferring situational meaning (ISM)	The ability to make inferences about details, relationships, situations, and arguments using textual or background knowledge.	<ul style="list-style-type: none"> - Identify and understand an implicit detail that is rewritten using different words. - Understand the underlying meaning of a sentence or a detail. - Understand the logical inference of an argument.
Understanding pragmatic meaning (UPM)	The ability to understand author's purpose, attitude, tone, mood, belief, and intention in the text.	<ul style="list-style-type: none"> - Understand the author's purpose, attitude, opinion, or stance on an issue in the text.. - Understand the general tone of a text. - Understand the purpose of the author via a detail in the text.
Lexical inferencing (LI)	The ability to guess the meaning of words using contextual clues.	<ul style="list-style-type: none"> - Guessing word meaning from contexts (words with different meanings) - Guessing word meanings from contexts (idiomatic expressions)
Identifying genre and text structure at discourse level (IS)	The ability to identify the genre (such as narrative, expository, persuasive, joke, diary) or identify structure of information and ideas at the discourse level (such as problem – solution, cause – effect, comparison, contrast).	<ul style="list-style-type: none"> - Identify the organizational structure of a text. - Identify the genre of a text.

5.2.2. Findings from the main stage

During the official judgment session, each expert was given the L-VSTEP reading test version A and the expert judgment form. However, instead of working collaboratively as in the pilot session, the experts independently answered the test questions, consulted the answer keys and conducted the judgment. After finishing the procedure, the two experts participated in a moderation session during which they discussed to reach agreement on the primary reading skill assessed by each item and other skills identified. Where disagreement occurred, they were required to offer justifications for their own decision, and if necessary, a new coding was added to the expert judgment form to reflect an additional skill judged to be assessed by the items.

The judgment practice was afforded by an expert judgment form developed from the test development guidelines, consensus on the interpretation of the skill descriptions in the expert judgment form during the pilot stage, and a moderation session where they discussed the judgment results to obtain agreement. Results of the expert judgment are presented in Table 5.2.

Table 5. 2. Results of the expert judgment of reading skills

Items	Primary skills identified	Potential involvement of other skills
Passage 1		
Item 1	Understanding explicit information	Understanding explicit information
Item 2	Integrating textual information	Integrating textual information, Inferring situational meaning, Understanding explicit information
Item 3	Lexical inferencing	Lexical inferencing, Understanding explicit information, Integrating textual information, Inferring situational meaning
Item 4	Inferring situational meaning	Understanding explicit information, Inferring situational meaning
Item 5	Understanding cohesive devices	Understanding explicit information, Understanding cohesive devices, Integrating textual information, Inferring situational meaning
Item 6	Lexical inferencing	Understanding explicit information, Lexical inferencing, Understanding cohesive devices, Integrating textual information, Inferring situational meaning
Item 7	Integrating textual information	Understanding explicit information, Integrating textual information

Item 8	Integrating textual information	Understanding explicit information, Integrating textual information, Inferring situational meaning
Item 9	Understanding cohesive devices	Understanding explicit information, Understanding cohesive devices
Item 10	Understanding pragmatic meaning	Understanding explicit information, Integrating textual information, Summarizing textual information, Understanding pragmatic meaning
Passage 2		
Item 11	Summarizing textual information,	Understanding explicit information, Summarizing textual information, Integrating textual information, Inferring situational meaning, Understanding cohesive devices,
Item 12	Understanding explicit information	Understanding explicit information
Item 13	Lexical inferencing	Understanding explicit information, Lexical inferencing, Integrating textual information, Inferring situational meaning
Item 14	Integrating textual information	Understanding explicit information, Integrating textual information, Lexical inferencing
Item 15	Understanding explicit information,	Understanding explicit information, Lexical inferencing
Item 16	Understanding explicit information	Understanding explicit information, Lexical inferencing
Item 17	Integrating textual information	Understanding explicit information, Understanding cohesive devices, Inferring situational meaning, Integrating textual information
Item 18	Inferring situational meaning	Understanding explicit information, Inferring situational meaning, Integrating textual information
Item 19	Understanding cohesive devices	Understanding explicit information, Understanding cohesive devices, Integrating textual information, Summarizing textual information
Item 20	Summarizing textual information	Understanding explicit information, Summarizing textual information, Inferring situational meaning, Integrating textual information
Passage 3		

Item 21	Inferring situational meaning	Understanding explicit information, Integrating textual information, Inferring situational meaning
Item 22	Understanding pragmatic meaning,	Understanding explicit information, Understanding pragmatic meaning, Inferring situational meaning, Integrating textual information
Item 23	Lexical inferencing	Understanding explicit information, Lexical inferencing, Integrating textual information, Inferring situational meaning
Item 24	Summarizing textual information	Understanding explicit information, Summarizing textual information, Integrating textual information
Item 25	Understanding cohesive devices	Understanding explicit information, Understanding cohesive devices, Inferring situational meaning, Integrating textual information
Item 26	Integrating textual information	Understanding explicit information, Integrating textual information, Understanding cohesive devices, Inferring situational meaning
Item 27	Understanding pragmatic meaning	Understanding explicit information, Understanding pragmatic meaning, Integrating textual information, Inferring situational meaning
Item 28	Lexical inferencing	Understanding explicit information, Lexical inferencing, Integrating textual information, Inferring situational meaning
Item 29	Understanding pragmatic meaning	Understanding explicit information, Understanding pragmatic meaning, Integrating textual information, Inferring situational meaning
Item 30	Understanding pragmatic meaning	Understanding explicit information, Understanding pragmatic meaning, Summarizing textual information, Integrating textual information, Inferring situational meaning
Passage 4		
Item 31	Understanding cohesive devices	Understanding explicit information, Understanding cohesive devices
Item 32	Understanding explicit information	Understanding explicit information
Item 33	Lexical inferencing	Understanding explicit information, Lexical inferencing, Inferring situational meaning

Item 34	Integrating textual information	Understanding explicit information, Integrating textual information
Item 35	Inferring situational meaning	Understanding explicit information, Inferring situational meaning, Integrating textual information
Item 36	Inferring situational meaning	Inferring situational meaning, Understanding explicit information, Integrating textual information
Item 37	Inferring situational meaning	Understanding explicit information, Integrating textual information, Inferring situational meaning
Item 38	Integrating textual information	Understanding explicit information, Integrating textual information
Item 39	Inferring situational meaning	Understanding explicit information, Inferring situational meaning, Integrating textual information
Item 40	Identifying text structure	Understanding explicit information, Identifying text structure, Integrating textual information, Summarizing textual information, Inferring situational meaning

As can be seen in Table 5.2, all the reading subskills specified in the test development guidelines were judged to be primarily assessed by at least one item, suggesting that the guidelines which informed the test development process was well-represented by the test items as judged by the experts. Integrating Textual Information was the most commonly assessed skill (primarily assessed by eight items), followed by Inferring Situational Meaning (seven items), and Lexical Inferencing (six items). Understanding Explicit Information, Understanding Cohesive Devices, and Understanding Pragmatic Meaning were each assessed by five items, while summarizing Textual Information and Identifying Text Structure were targeted by only three items and one item respectively. Except for Item 1, Item 12, and Item 32, all other items require a combination of at least two subskills to arrive at the answer. Item 11 and Item 30 were judged by the experts to require readers to engage with a maximum of five subskills included in the guidelines, while the majority of other items demand three or four subskills. Understanding Explicit Information at Local Level should be employed in responding to all the items in the test, while Integrating Textual Information appears in the process of answering a majority of other items (29 items). Identifying Text Structure should be exercised to answer only one item, suggesting that the ability to use this skill is not necessary to do well on the test overall.

5.2.3. Discussion of the expert judgment findings

Notable skill use patterns emerged from the findings presented above. First, the experts identified all eight reading skills included in the test development guidelines, of which only one skill – “Understanding Explicit Information” was considered a lower-level process of reading, as defined by Grabe (2009); Khalifa and Weir (2009). This is reasonable given that the test is developed to assess reading proficiency at levels three (intermediate) to five (advanced) according to the CEFR-VN. Second, in line with contemporary perspectives of L2 reading processes and empirical evidence from L2 reading assessment research (Nassaji, 2003; Rupp et al., 2006), the process of answering a specific test item in this study was believed by experts to involve multiple subskills both at lower and higher levels. Out of the 40 test items, 37 items were judged to require at least two reading subskills. In addition, the skill “Understanding Explicit Information” – a lower-level process of reading was judged to be involved in answering all 40 items, resulting in the potential combination of at least one lower-level process and one higher-level process of reading to answer the test items. This reflects the interactive perspective of L2 reading which is deemed to entail the integration of multiple components, both at lower-level text-based and higher-level reader-based processes (Nassaji, 2003). The identification of “Understanding Explicit Information” in answering all 40 items also implies that lower-level processes play an essential role in the reading process, the absence or ineffectiveness of which affects the execution of higher-level processes and ultimately impacts negatively on overall comprehension.

All items believed by experts to primarily assess “Lexical Inferencing” (Items 3, 6, 13, 23, 28, 33) ask test-takers to identify a word closest in meaning to a given word in the passage. This type of question was thought by experts to involve the use of “Understanding Explicit Information”, “Inferring Situational Meaning” and “Lexical Inferencing”. In five items, “Integrating Textual Information” was required as readers needed to synthesize information across sentences for the inference making process, while in one item, the information needed to guess the word meaning was contained in a single sentence. “Understanding Cohesive Devices” was required in one item since the given word was preceded by a referent word (their) which required readers to identify its antecedent.

“Inferring Situational Meaning” was believed to be primarily assessed by Items 4, 18, 21, 35, 36, 37, and 39, all requiring readers to employ “Understanding Explicit Meaning” and “Integrating Textual Information”. This is theoretically logical because readers needed to connect

pieces of information across sentences to enrich a text model of comprehension before building a situation model of interpretation where they seek to understand not only the message conveyed by the text, but also the inferences beyond it (Grabe, 2009).

“Understanding Pragmatic Meaning” is the primary skill targeted by Items 10, 22, 27, 29, and 30. Answering this type of question required the potential involvement of a range of different subskills, including “Understanding Explicit Information”, “Integrating Textual Information”, “Inferring Situational Meaning”, and in some cases “Summarizing Textual Information”. Khalifa and Weir (2009) believed that pragmatic inference questions are difficult because readers have very different knowledge, experience and opinions, so that given the same information in a text, different readers process it with very different perspectives and expectations. Therefore, the deployment of a wide range of subskills in answering this type of question seems to compensate for the inherent complexity associated with it.

Except for a few items (Items 1, 9, 12, 15, 16, 31, and 32), which assessed readers’ understanding at the local level only, the majority of items in the test were believed by experts to require understanding at the global level and the ability to combine multiple subskills in order to arrive at the answers. This is in line with the reading proficiency levels targeted by the test as it is assumed that candidates taking the test should have achieved reading proficiency at levels one and two in the CEFR-VN which are described respectively as the ability to recognize familiar lexical items and understand simple sentences; as well as the ability to understand basic information and short, simple texts related to areas of immediate relevance.

The expert judgment results revealed the reading subskills that are primarily tested by the items as well as the potential involvement of other skills during the reading processes. However, the identification of these skills only reflects the experts’ views about the reading processes informed by the test development guidelines. It is, however, also essential to understand and document the reading processes engaged by the test-takers during the test, and whether these processes align with what is expected by the experts. Exploring the alignment between the reading processes expected by the experts and those reported by the test-takers provides important evidence for the validity of the interpretation and use of the L-VSTEP reading test. The next section, therefore, is devoted to this essential aspect of the validation process.

5.3. Findings from students’ verbal reports

Since the primary purpose of this phase was to examine the reading subskills employed by the test-takers in the reading process and the extent to which those skills align with the experts' judgment, the findings were presented according to the primary subskill deemed by experts to be assessed by each item and the reading process of the test-takers at different proficiency levels when answering that item.

5.3.1. Items mainly assessing Understanding Explicit Information

The subskill Understanding Explicit Information was believed by experts to be mainly assessed by Items 1, 12, 15, 16, and 32. All of the items asked about a specific detail in the text and required only understanding of information within a sentence to find the answer. As indicated in Table 5.3 below, the majority of the students across proficiency levels reported employing Understanding Explicit Information to answer these items. In a few cases, Syntactic Parsing, Lexical Access, and Eliminating Implausible answers were also utilized. For example, seven students across three proficiency levels reported to have used Syntactic Parsing to answer Item 15 the correct option of which is a paraphrase sentence of another sentence in the text. This process is illustrated by Student 1: “... *I read the sentence on line 16, and recognized that option B is actually a paraphrased version of the sentence on line 16, ‘expectations are very high’ is the same as ‘renters over-expect’...*”

Table 5. 3. Findings from students' protocols on the Understanding Explicit Information subskill

	Item 1			Item 12			Item 15			Item 16			Item 32			
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	
Using key words in questions to locate information in the text (KM)	1/3			1/3												
Lexical access (LA)										3/3		1/3				
Syntactic parsing (SP)							2/3		2/3		3/3					
Understanding information within a sentence (UEI)	3/3	3/3	3/3	3/3	3/3	3/3	3/3	3/3	3/3	3/3	2/3	1/3	1/3	3/3	3/3	3/3
Understanding cohesive devices (UCD)																
Integrating information across sentences (ITI)							1/3									
Understanding main ideas and supporting details (STI)																
Inferring situational meaning (ISM)							1/3									
Lexical inferencing (LI)							1/3									
Understanding pragmatic meaning (UPM)																
Recognizing text structure (RTS)																
Eliminating implausible answers (EIA)	1/3		1/3					1/3								

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

Lexical Access, retrieving lexical knowledge from mental lexicon, was used to answer Item 16 by three high-proficiency students and one middle proficiency student. This item asked about a detail in the text, but the answer could be found if the students knew the meaning of the key word “grudgingly” given in the key sentence in the text the synonym of which - “reluctant” - is the correct option. All four students reported that they found the answer because they knew the meaning of the key word in the text that matched with the correct option. In addition, two low-proficiency students could not answer the question because they did not have knowledge of either the word in the text or the given word in the option.

Student 2: “... I chose C because I found the word ‘grudgingly’ in paragraph 5. I know this word. It means unwilling or reluctant ... I also found ‘raise the price limit’ (immediately after the word ‘grudgingly’ in the passage). It refers to ‘renters’, so I chose reluctant.

Student 9: ... I don’t know the meaning of ‘grudgingly’, so I don’t understand the attitude of the ‘renter’. I chose the answer randomly.”.

Eliminating Implausible Answers was reported by two high-proficiency students and one mid-proficiency student when they answered Item 1 and Item 15. For example, although the answer to Item 15 could be found by using Understanding Explicit Information the process of answering this item by one high-proficiency student was rather complicated so that rather than using Understanding Explicit Information, she used a combination of Integrating Textual Information, Inferring Situational Meaning, and Eliminating Implausible Answers. The student still arrived at the correct answer but showed a deep understanding of the text and the ability to process a large chunk of message rather than focusing specifically on a single sentence. The quote below helps clarify this process.

Student 2: “In paragraph 7, there is a sentence ... ‘it is not uncommon in New York to ... only to find out’ ..., which means that they want to find an apartment but end up finding a small one with only a wall ... they have high expectations when they rent an apartment but the outcome makes them surprised. I don’t find information for other options.”.

In a nutshell, the five items deemed by experts to primarily assess Understanding Implicit Information have induced the participants across proficiency levels to consistently use this subskill to find the answers. The analysis of the students’ verbal protocols also revealed the use of Syntactic Parsing (SP) and Lexical Access (LA) to answer Item 15 and 16 respectively. All three identifiable

skills from students' protocols are considered lower level processes of reading and contribute to readers' understanding at the local level where they need to recognize the words, process the sentence structures and retrieve explicit meaning from the sentences. These subskills were not identified by the experts probably because they are not included in the test guidelines, and experts approached the judgment task with the belief that potential test candidates are assumed to have these foundational word recognition and syntactic parsing subskills. In addition to the subskills involved in the reading process, the verbal protocols also introduced the test-taking strategy of Eliminating Implausible Answers. This strategy, however, was reported by only one high-proficiency and one mid-proficiency level student who used it for confirmation of their text understanding.

5.3.2. Items mainly assessing Lexical Inferencing

The expert judgment process yielded six items (Items 3, 6, 13, 23, 28, 33) which were thought to mainly assess Lexical Inferencing subskill. Other subskills including Integrating Textual Information, Inferring Situational Meaning, and Understanding Explicit Information were also believed to be involved in the reading process. Furthermore, the experts unanimously agreed that test-takers needed to have knowledge of either the given words or the words in the texts to answer all the questions. Results of the verbal protocol analysis are presented in Table 5.4.

Table 5. 4. Findings from students' verbal protocols on the Lexical inferencing subskill

	Item 3			Item 6			Item 13			Item 23			Item 28			Item 33		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
Using key words in questions to locate information in the text (KM)																		
Lexical access (LA)	3/3	3/3	3/3	3/3	2/3	1/3	3/3	3/3	1/3	3/3	3/3	2/3	3/3	3/3	2/3	3/3	3/3	3/3
Syntactic parsing (SP)																		
Understanding information within a sentence (UEI)	2/3	2/3	2/3	1/3		2/3			2/3	1/3	2/3		3/3	3/3	3/3		1/3	1/3
Understanding cohesive devices (UCD)	1/3																	
Integrating information across sentences (ITI)	2/3	1/3					1/3			1/3								
Understanding main ideas and supporting details (STI)																		
Situational Inferencing (SI)	3/3	1/3	1/3	2/3	1/3		1/3	1/3		1/3	2/3		2/3		3/3	2/3	1/3	
Lexical inferencing (LI)	3/3	1/3	2/3	2/3	1/3	1/3	1/3			1/3	1/3		2/3	3/3	3/3	3/3	2/3	3/3
Understanding pragmatic meaning (UPM)																		

Recognizing text structure (RTS)								
Eliminating implausible answers (EIA)	1/3		1/3	1/3	2/3	1/3	1/3	1/3
Replacing for confirmation (RFC)			1/3	1/3				
Uninformed guessing (UG)				1/3				

+ *The number before the slash indicates the number of students in each proficiency group who answered the item correctly.*

+ *High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)*

Findings from students' verbal protocols suggested that students used the majority of subskills that were also identified by experts. All protocollers reported to have used Lexical Access, Understanding Explicit Information (UEI), Understanding Situational Meaning (ISM), and Lexical Inferencing (LI) to answer the majority of the items under examination. However, Integrating Textual Information, which was believed by experts to be involved in answering all 6 items, turned out to be the least frequently used subskill by the students. Only two high-proficiency students and one middle proficiency student employed this subskill in the reading process to answer Items 3, 13 and 23. One salient pattern of skill use emerging from students' verbal reports was the combination of Lexical Access, Understanding Explicit Information, Inferring Situational Meaning, and Lexical Inferencing. This is understandable, given that readers need to retrieve word meaning from their mental lexicon, understand the sentence in which the word occurs, make inferences based on the clues contained in the sentence and match the word with the alternatives. One typical example is Item 28 which requires test-takers to find the word that can replace the word "scorches" in the reading passage. Some students could use the clues in the immediate sentences such as "the sun", "the field", and "heat" to make an inference about the word meaning and found the correct answer, which is "burns". For example, student 5 reported that *"I don't know the meaning of 'scorches', I think the sun can only 'burns', the sun's heat can only burn the fields, not warms up or shines, so 'burns' must be the answer."*

Where clues in the given sentence were not clear or it was hard for them to connect different clues, they relied solely on their vocabulary knowledge to tackle the item. For example, student 1 reported that *"I know the meaning of 'scorches'. Initially, I thought A, B, and C are all logical because they all mean making something warm, but finally I think 'burns' has the strongest nuance of meaning and should be more suitable."*

Two students in the middle proficiency group followed the same logical reasoning as student 5 but ended up choosing "heats up". This could be explained by their incorrect inference of the meaning of the phrasal verb "heats up".

Student 4: "... from the context, I saw 'the sun', 'the field', so I think 'the sun' makes 'the field' hot, then I saw 'raising ...', so I think 'heats up'. 'heats up' – 'up' means go up, ... higher ... so I think it has the strongest meaning."

The use of Integrating Textual Information was minimal in contrast to what was expected by the experts. This could be because the clues in the sentences that contained the given words

were rich enough for the participants to garner and make lexical inferences. This was further aided by the task type – multiple choice items – where students can weigh the alternatives against each other to make decisions. The participants at middle and low proficiency levels employed Eliminating Implausible Answers to answer all the items in this section except for Items 3 and 6. Only one student from the high proficiency group used Eliminating Implausible Answers, but mainly for the purpose of confirming his lexical inferencing. All three high-proficiency students could answer the six items correctly by either drawing on their vocabulary knowledge or making inferences based on the sentential clues. One student even combined different skills and strategies including Lexical Access, Lexical Inferencing, Integrating Textual Information, Inferring Situational Meaning, and Eliminating Implausible Answers to answer Item 13, while another high proficiency student used the strategy “Replacing for Confirmation” – putting the option word in the text and reinterpreting the sentence to retrieve meaning to answer Item 3, 13, and 28. Low-proficiency students answered four of the six items incorrectly, either because they did not have knowledge of the word or retrieved the wrong meaning of the given words. In a few cases, they made guesses due to a lack of knowledge of the word in the question stem as well as in the options.

5.3.3. Items mainly assessing Understanding Cohesive Devices

The experts identified five items (5, 9, 19, 25, and 31) that mainly assess the Understanding Cohesive Devices (UCD) subskill. Among them, Items 9 and 31 could be answered by using information and clues within the immediate sentences that contained the referent words. The reading process was therefore believed to involve the use of only two subskills, Understanding Explicit Information (UEI) and Understanding Cohesive Devices, where readers needed to understand the sentence and identified the antecedent of the given referent word. On the other hand, Items 5, 19 and 25 were thought to require the use of Integrating Information Across Sentences (ITI) and Inferring Situational Meaning (ISM) in addition to Understanding Explicit Information and Understanding Cohesive Devices because the answers to these items could not be leaned on the information within a single sentence. Findings from students’ verbal protocols are presented in Table 5.5.

Table 5. 5. Findings from students' verbal protocols on the Understanding Cohesive Devices subskill

	Item 5			Item 9			Item 19			Item 25			Item 31		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
Using key words in questions to locate information in the text (KM)															
Lexical access (LA)															
Syntactic parsing (SP)															
Understanding information within a sentence (UEI)	3/3	3/3	3/3	3/3	3/3				3/3		1/3	3/3	3/3	3/3	3/3
Understanding cohesive devices (UCD)	3/3	3/3	1/3	3/3	3/3		3/3	3/3	3/3	3/3	3/3		3/3	3/3	3/3
Integrating information across sentences (ITI)	3/3	3/3	1/3				3/3	3/3	3/3	3/3	3/3	1/3			
Understanding main ideas and supporting details (STI)															
Situational Inferencing (SI)	2/3						2/3			2/3		1/3			
Lexical inferencing (LI)										1/3					
Understanding pragmatic meaning (UPM)															
Recognizing text structure (RTS)															

Eliminating implausible answers (EIA)	1/3	1/3	1/3	1/3
Uninformed guessing (UG)	2/3			
Replacing for confirmation	1/3			1/3

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

In general, all the subskills identified by the experts were employed by the participants in answering the questions. The skill use pattern was also fairly well aligned with what was expected by the experts. Understanding Cohesive Devices and Understanding Explicit Information were reported by all participants to answer Item 9 and Item 31 except for two low-proficiency students who admitted to making blind guesses about item Item 9. Understanding Explicit Information, Understanding Cohesive Devices, and Integrating Textual Information were employed by the majority of the participants to answer Item 5, Item 19 and Item 25. However, Understanding Explicit Information was not identifiable from high- and mid- proficiency students' reports in answering Item 19 and Item 25. Item 19 asks about the antecedent of the word "those" which is placed at the beginning of a new paragraph. It is therefore assumed that readers need to integrate information in the previous paragraphs to find the answer. High- and mid- proficiency students mainly focused on reporting the process of connecting information across sentences to make reference to the given word rather than explaining their understanding of a single sentence.

Student 2: "The previous paragraph mentioned both graduates and landlords, and the amount of money that graduates need to make to satisfy the landlords, so it must be graduates."

Though not explicitly stated by the protocollers, it is conceivable that they also employed Understanding Explicit Information in the reading process because the integration of information across sentences require the retrieval of meaning from single sentences. One low-proficiency student, though, employed all expected subskills as did other students, and chose "landlords" rather than "graduate" – the correct answer for Item 19.

Student 8: "... I saw that 'landlords' were repeatedly mentioned in the previous paragraph such as in line 15, 23, 25, ... in the below sentence, the landlords required students to have parents, so I chose landlord..."

It can be inferred from this student's protocol that although he located and combined information beyond a single sentence to identify the antecedent of the word "those" his use of these skills were inefficient since the word "graduates" was also mentioned repeatedly in the previous sentences in the text and exhibited a clearer connection to the word "those". This inefficiency in skill use was also indicated in his interpretation of the sentence that contains the referent word. Instead of interpreting the original sentence in the text as "students, who don't make

enough money to pay the rent, need a guarantor”, this student wrongly interpreted the same idea as “landlords who don’t make enough money, need a guarantor”.

Inferring Situational Meaning (ISM) – a core skill thought by experts to contribute to answering the items in the table above was only reported by two students in the high-proficiency groups. One low-proficiency student also used this subskill to answer Item 25. This could be explained by the fact that Inferring Situational Meaning can be subsumed in the process of understanding reference, and therefore became less salient in the students’ reports. In fact, Khalifa and Weir (2009) considered recognizing “anaphoric reference” as a component of the inferencing process which also entailed lexical inferencing and pragmatic inferencing. That high-proficiency students could verbalize in detail the process of making situational inference can be attributed to their good working memory and the deeply engaging process of deciphering and integrating textual information, which made the thought processes readily available for access during their reporting. An example can be found in Student 2’s report.

“The author used the word ‘broken’ which means cannot be the same, ... ‘Obama is supportive’, so other politicians cannot be supportive. Because they are not supportive, they take no action, so I chose C.”.

In addition to the core subskills involved in answering the items, students also employed three strategies of Eliminating Implausible Answers (EIA), Replacing for Confirmation (RFC) and Uninformed Guessing (UG), though only occasionally. Replacing for Confirmation and Eliminating Implausible Answers were used by high- and mid- proficiency students as a method of carefully checking their answer, while Uninformed Guessing was used by students toward the lower end of the proficiency spectrum when they could not find the information in the text.

5.3.4. Items mainly assessing Integrating Textual Information

Eight items (2, 7, 8, 14, 17, 26, 34, and 38) were identified by experts as mainly testing the Integrating Textual Information (ITI) subskill. Items 14, 34, and 38 ask test-takers to choose the statement that is not mentioned in the texts. In order to answer this question type, it was believed that students should be able to retrieve meaning from a single sentence (UEI) and connect information across sentences (ITI) to identify the statements that are not supported by the text. Two items, 17 and 26, require test-takers to put a given sentence in the correct place in the relevant reading texts. The experts believed that students should be able to understand information within and across sentences (Understanding Explicit Information and Integrate Textual Information),

Understand Cohesive Devices (UCD), as well as to make Situational Inference (ISM) to find the correct place for the given statement. All the aforementioned items require understanding at the cross-paragraph level to yield the needed information, while the other three items (Items 2, 7, and 8) can be answered by using the information within paragraphs. Table 5.6 summarizes findings from the students' verbal protocols.

Table 5. 6. Findings from students' verbal protocols on the Integrating Textual Information subskill

	Item 2			Item 7			Item 8			Item 14			Item 17			Item 26			Item 34			Item 38			
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	
KM				1/3									1/3												
LA																									
SP																									
UEI	2/3			2/3		3/3	3/3	3/3	2/3	2/3	2/3	3/3	2/3	3/3	1/3	1/3	2/3	2/3		1/3	2/3	3/3		3/3	3/3
UCD														3/3	2/3	1/3	3/3	2/3							
ITI	3/3	3/3	1/3	2/3			3/3	1/3		3/3	3/3		3/3	3/3		3/3	3/3		3/3	3/3	1/3	3/3	3/3		
STI																									
SI	3/3	3/3	1/3			1/3	3/3	3/3					3/3	1/3		3/3	2/3			1/3				1/3	
LI																									
UPM																									
RTS																									
EIA	1/3	1/3								3/3	3/3					1/3		3/3	3/3	1/3	3/3	3/3			
TTE				1/3												1/3									
RFC													1/3	2/3											
UG																	2/3							2/3	

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

As expected by the experts, Understanding Explicit Information and Integrating Textual Information were invoked during the students' reading processes to answer Items 14, 34, and 38. All three low-proficiency students answered Item 38 incorrectly due to Uninformed Guessing and making incorrect inferences. In the case of Item 14, these students also chose the wrong answer as they misunderstood a detail mentioned in a single sentence in the text without considering details in other options and sentences. As reported by one student: *"I think 'to look in other neighbourhood or get a roommate' means to find a roommate who is living in a nearby neighbourhood, so I choose D – 'they decide to look for a place in a different neighbourhood' – as not mentioned in the text."* An additional strategy employed by all the high- and mid-proficiency students to answer Items 14, 34, and 38 was Eliminating Implausible Answers (EIA). One high-proficiency student reported: *"I chose B because this question ... we choose the not-mentioned sentence, I first find the mentioned sentences, when I found them, I put a tick at the end of the sentences for elimination, the sentence left is the correct one."* Rather than using Eliminating Implausible Answers as a strategy to confirm the answers, or to weigh different alternatives against the information in the text, Eliminating Implausible Answers is used here as an essential strategy in the reading process to answer this type of question.

The four subskills that were used by the majority of students to answer Items 17 and 26 were Understanding Explicit Information, Integrating Textual Information, Inferring Situational Meaning and Understanding Cohesive Devices. The low-proficiency students seem to be able to use only Understanding Explicit Information and, in a rare case, Understanding Cohesive Devices to find the answers to these items. This reflects their limited ability to simultaneously employ a range of different skills – those at the higher levels of reading processing - to make sense of the text. It is evident from the verbal protocols that in addition to Understanding explicit information and Integrating textual information, the ability to identify the cohesive devices such as discourse markers (Item 17) and repeated words (Item 26) plays an important role in assisting readers to appreciate the cohesion of the text, which in turn helps them to establish the connection between the given statement and the rest of the paragraphs. The following quotes help illustrate this point.

Student 1: "First, I look for the linguistic signals in the given sentence. Here I found the word 'aside from', and 'price' and 'limit'. It means that 'price' and 'limit' and something else. So, I think it must follow the parts where they discuss 'price and limit'. I chose D because space was described as small, and price, here in this part, '... first shock'."

Student 4: “I chose D because I saw ‘to start with’, so I think there should be a sentence before that. ‘To start with’ is to give an example, or an illustration for something else. I put D here and reinterpret the paragraph and I found it logical.”.

Student 2: “I chose D because I can see the country name here. This part talks a lot about countries, ‘China, Greenland, India’, ... so I think the connection of ideas here is about countries.”.

In addition to the core reading skills employed by the students, two additional strategies of Replacing for confirmation (RFC) and Test Taking Experience (TTE) were also reported by three students in the high- and mid- proficiency groups and one student in the mid-proficiency group respectively. The former was used to reconfirm if the chosen statement fitted meaningfully into the paragraphs while the latter was employed by the student to locate the necessary information in the text where the answer could be found. This is illustrated in the following excerpt:

Student 4: “I chose D because once again, I think the answer to this question must follow the flow of reading so far. You cannot go back to A and B to find the answer. I think C and D are the correct answers. This sentence starts with ‘India’, I also found ‘China, Greenland, Ireland, Antarctica’, ... so I think they are listing the country names.”.

Understanding Explicit Information, Integrating Textual Information and Inferring Situational Meaning were employed by almost all students to answer Items 2 and 8, which also reflects experts’ judgment. The students, however, seemed to be able to answer Item 7 by drawing on their interpretation of a single sentence. For example, student 5 reported: *“Whose job involves in a large part listening to others. I chose D because I found this detail, ‘have to remember a huge task of what you do is listening.’* Only two high-proficiency students integrated information across the paragraph to find more support for this statement. For example, Student 1 recounted:

“In line 2, ‘have to remember a huge task of what you do is listening’. Listening here means listening to different people, I read the rest of the paragraph and found that these different people were mentioned, such as ‘advocates, witnesses, defenders’. I also read other paragraphs to find support for the statement, but I did not find any information. So I chose D.”.

5.3.5. Items assessing Inferring Situational Meaning

Seven items including Items 4, 8, 21, 35, 36, 37, and 39 were judged by experts to mainly test readers’ ability to make inference based on the information in the text and their background

knowledge (Inferring Situational Meaning, ISM). Understanding Explicit Meaning (UEI), Integrating Textual Information (ITI) and Inferring Situational Meaning (ISM) were thought to contribute to deciphering the answers to all these items. Item 4 also potentially requires the use of Understanding Cohesive Devices (UCD) since students need to unlock the antecedent of a referent word to understand a key sentence in the reading process.

Table 5. 7. Findings from students’ verbal protocols on the Inferring Situational Meaning subskill

	Item 4			Item 18			Item 21			Item 35			Item 36			Item 37			Item 39		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
KM	3/3	1/3					1/3						1/3			1/3			1/3		
LA																					
SP																					
UEI	1/3	1/3	3/3	1/3	2/3	2/3	2/3	1/3	3/3	1/3			3/3	2/3	2/3	3/3	1/3	3/3	2/3	1/3	2/3
UCD																					
ITI				3/3	3/3		2/3	1/3		1/3	1/3		2/3	1/3		2/3	1/3		3/3	2/3	2/3
STI																					
SI		1/3		3/3	3/3		3/3	3/3	3/3	1/3	2/3		3/3	3/3		2/3	1/3		1/3	1/3	1/3
LI				1/3			3/3	1/3			1/3		1/3								
UPM																					
RTS																					
EIA	2/3			2/3	2/3		2/3	2/3		2/3											
TTE																					
RFC																					
UG						1/3				1/3	1/3	3/3			1/3						

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

As presented in Table 5.7 above, students' verbal protocols revealed relatively divergent patterns of skill use as compared with expert judgment. The use of Inferring Situational Meaning – the primary skill believed to be assessed by these items also differed across items and proficiency levels. Item 4 which asks about readers' interpretation of an opinion in the text generated an unexpected pattern of skill use. Except for one mid-proficiency student who reported the use of Inferring situational meaning, all other students relied on Keyword Matching (KM), Eliminating Implausible Answers, and Understanding Explicit Information to find the answer. The use of Understanding Cohesive Devices – another important subskill in the reading process as expected by the experts - was completely absent from students' reports. All nine students answered the item incorrectly. The following excerpts might help unpack the underlying reasons.

Student 1: "... limitations take years, I use the elimination methods, 'limitation' here refers to weakness, 'learning your limitation' means controlling your weakness, so I chose A."

Student 5: "I chose A because I can see the synonyms here, 'limitation' versus 'weakness', 'learning' versus 'controlling', and 'take years' versus 'take a long time'. I see these two sentences are paraphrases of each other, so I did not read the text."

It seems evident from students' responses that they relied on matching keywords in the question stem with those in the options without much reference to the text. They missed an important piece of information in the text which was referred to via the cohesive device of anaphoric reference, so that what takes time is not the learning of limitations but the learning of how to perform a brain operation. Therefore, it can be inferred from these verbal protocols that, this item has failed to assess their ability to make situational inference and to engage them in processing the text.

All students other than those in the low-proficiency group reported to have employed the core reading skills suggested by experts to answer Items 18, 21, 35, 36, and 37. In a few cases, additional strategies/subskills of Eliminating Implausible Answers, Lexical Inferencing and Keyword Matching were also reported. For instance, one mid-proficiency student reported using Lexical Inferencing in answering Item 35:

'Incredibly lucky' means too lucky, the chance of happening is very small, it means nearly impossible, so I chose the answer because of this word, I didn't read other sentences".
(Student 4).

The other student reported using Keyword Matching and Inferring Situational Meaning to answer item 21: *“I chose A, ‘unusual’, it’s the opposite of ‘habitual’. I saw the sentence ‘instead of the habitual snowing landscape’, ... so unusual is the correct answer.”* (Student 3).

Low-proficiency students reported mainly retrieving information from a single detail or sentence in the text and sometimes made inferences from these details. This process of reading proved ineffective as they answered the majority of the items incorrectly in this part. For example, Student 7 reported: *“I chose C because I saw in line three the sentence ‘in my 15 years of flying I have not seen a scene like this’, so I think this scene must be worthy for him.”* (incorrect inference). Another student reported: *“I chose D because in the first paragraph, the author said that when he opened the window, he saw a large sky, and a spacecraft, so I think it must be a very magnificent scene.”* (Student 9, incorrect understanding of details).

In sum, except for Item 4 which seems to have misled readers and Item 39 which functioned as expected by the experts, the other items elicited a range of reading subskills and strategies to varying degrees by the protocollers. This could be explained by the fact that making an inference is a higher-level process of reading (Weir & Khalifa, 2009), and each reader answers the test items with different background knowledge and different interpretation of the text (Alderson, 2000). The extent to which readers employ the reading skills to answer this type of question, therefore, varies with respect to their level of text understanding and background knowledge. Although these questions elicited a wide range of reading subskills to varying degrees, they have functioned well in terms of discriminating readers at different proficiency levels. Those to the lower end of the proficiency spectrum used limited subskills and primarily rely on understanding at the sentential level. On the other hand, those at the higher end employed different reading subskills and strategies in the reading process, most of which accord with experts’ expectations.

5.3.6. Items assessing Understanding Pragmatic Meaning

Items 10, 22, 27, 29 and 30 were judged by experts to mainly assess Understanding Pragmatic Meaning (UPM) subskill. All these items ask about the purpose and attitude of a detail in the text or of the author. In order to answer these items, the experts unanimously agreed that students should employ a range of different subskills including Understanding Explicit Meaning (UEI), Integrating Textual Information (ITI), Inferring Situational Meaning (ISM), and Understanding Pragmatic Meaning. In addition, Items 10 and 30 also require the use of Summarizing Textual Information (STI). Although the items ask about the purpose, attitude and

tone of the author or a detail in the texts, they all can be answered by using information within the texts rather than using readers' pragmatic knowledge only.

Findings from students' verbal protocols are presented in Table 5.8 below. Integrating Textual Information and Inferring Situational Meaning were employed by students to answer all the items under examination, though the use pattern varied across proficiency levels. While these two subskills were employed by the majority of high- and mid- proficiency level students to answer all the items, they are only identifiable from low-proficiency students' verbal reports in answering Items 10 and 30. Summarizing Textual Information was reported by high- and mid- proficiency students when they answered Items 10 and 30, as expected by experts. However, the use of this subskill is absent from low-proficiency level students' reports. Similarly, Understanding Pragmatic Meaning, the primary subskill assessed by the items was only reported by students at the higher end of the proficiency spectrum. These findings imply that lower proficiency students differ from higher proficiency students in terms of their limited skill range and inappropriate skill use, and that these items have functioned well in terms of discriminating students at different proficiency levels. The majority of protocollers reported using the information in the text to answer the items instead of using their pragmatic knowledge as illustrated in the following excerpts.

Student 4: "I think that the passage is about different people, with different jobs, so I think the purpose is to report what different people do and think. I spend a few minutes skimming through the passage to get its gist."

Student 2: "I think that the author is worried, he is very much concerned about other people. The author talks about the solutions, the impacts they have. He is also shocked when he saw the scene. So, I think his attitude is supportive."

Student 1: "... so, this question, I have to make inference based on the tone of the author. For example, he said '...of course ...', which means something sympathetic. Although he took lots of actions, he encountered obstacles from others..."

Table 5. 8. Findings from students' verbal protocols on the Understanding Pragmatic Meaning subskill

	Item 10			Item 22			Item 27			Item 29			Item 30		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
KM															
LA															
SP															
UEI					1/3	2/3		1/3			1/3				
UCD															
ITI	3/3	2/3	3/3	2/3	2/3		3/3	1/3		2/3	1/3		3/3	3/3	1/3
STI	3/3	3/3		2/3						2/3			1/3		
SI	2/3	2/3	1/3	3/3	2/3		3/3	1/3		3/3	2/3		3/3	3/3	2/3
LI											2/3				
UPM				1/3	1/3		3/3	2/3		2/3	2/3				
RTS															
EIA	2/3							1/3			1/3				
TTE															
RFC															
UG						1/3		1/3			1/3	2/3			1/3

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

Although Understanding Pragmatic Meaning is the primary subskill assessed by the items as expected by experts, the use of this subskill was not clearly identifiable from students' reports. What was more pronounced via students' protocols was the process of retrieving literal meaning from individual sentences and making inference based on those sentences. Minimal pragmatic inference was made as students did not use their pragmatic knowledge in the reading process. This is evident from a student's excerpt,

Student 2: *"I chose D, I am not very good at questions about attitude. Therefore, I first translate the options into Vietnamese, then I see which one is more appropriate. I ponder B and D, and I chose D because I didn't see where the author said he was appreciative. I don't use pragmatic inference because it is influenced by culture and language use. In Vietnamese, that phrase may mean 'supportive', but in English it may have a different meaning. I am not sure. So, I have to find evidence from the passage to support it."*

5.3.7. Items mainly assessing Summarizing Textual Information

Items 11, 20, and 24 were judged to mainly test Summarizing Textual Information (STI) subskill. Answering these items potentially involves the use of multiple subskills in the reading process. Item 11 seems to elicit the widest range of subskills of all the items in the test, including Understanding Explicit Meaning (UEI), Integrating Textual Information (ITI), Inferring Situational Meaning (ISM), Summarizing Textual Information (STI), Understanding Cohesive Devices (UCD), and Recognizing Text Structure (RTS). Items 20 and 24 were thought to require the use of Understanding Explicit Information, Summarizing Textual Information, Integrating Textual Information, and Inferring Situational Meaning.

Table 5. 9. Findings from students’ verbal protocols on the Summarizing Textual Information subskill

	Item 11			Item 20			Item 24		
	H	M	L	H	M	L	H	M	L
Using key words in questions to locate information in the text (KM)	3/3		1/3	3/3	1/3				
Lexical access (LA)									
Syntactic parsing (SP)									
Understanding information within a sentence (UEI)				2/3	1/3	3/3	2/3	1/3	3/3
Understanding cohesive devices (UCD)	3/3		2/3		1/3				
Integrating information across sentences (ITI)	2/3	2/3	1/3	3/3	1/3		3/3	2/3	2/3
Understanding main ideas and supporting details (STI)	1/3	1/3		2/3			1/3	2/3	
Situational Inferencing (SI)		2/3	1/3	3/3	3/3	3/3	2/3		1/3
Lexical inferencing (LI)									
Understanding pragmatic meaning (UPM)									
Recognizing text structure (RTS)							1/3	1/3	
Eliminating implausible answers (EIA)	2/3			2/3	2/3			1/3	
Uninformed guessing (UG)							1/3		
Replacing for confirmation									
Test taking experience (TTE)				1/3					

+ The number before the slash indicates the number of students in each proficiency group who answered the item correctly.

+High proficiency students (H); Middle proficiency students (M); Low proficiency students (L)

As presented in Table 5.9, aside from the core reading subskills suggested by the experts, another important strategy that emerged from the students' reading processes is Keyword Matching (KM). This strategy together with Understanding Cohesive Devices and Integrating Textual Information were employed by all high-proficiency students to correctly answer Item 11. One student reported:

“Question 11 is about the title of the passage. I read all the paragraphs. I found it easy to follow the passages because each passage starts with a cohesive device: ‘first shock’, ‘second shock’. ‘Shock’ means surprises, so I chose C, surprises that await first-time renters.” (Student 1).

Similar reading processes were reported by another high-proficiency student:

“I underline the keywords in the question, surprising and first-time renters. In the passage I found surprising and newcomers. Each paragraph starts with ‘first shock’, ‘second shock’, the same meaning as surprise, so I chose C.” (Student 3).

One low-proficiency student could answer this item correctly, but rather than combining Keyword Matching, Understanding Cohesive Devices, and Integrating Textual Information, he primarily relied upon Understanding Cohesive Devices and Keyword Matching: *“I chose C because I saw the words ‘first shock’ and ‘second shock’, ... two times of shock, similar to surprise (Student 9).”*. Surprisingly, all three students in the mid-proficiency group answered this item incorrectly due to either making a wrong inference as reported by student 4 or forgetting to come back to the item after answering all other items as reported by student 5.

Student 4: *“After the reading the passage, I think apartments in New York are expensive, but the demand of the renters is very high, so I guess it is not sure whether they can buy an apartment in New York.”*

Student 5: *“This is the question about the best title, so I must do it after I read the passage and answer other questions. Then I forget to answer this item.”*

Item 20 rather than Item 11 elicited the widest range of reading subskills as reported in students' protocols. Core reading subskills of Understanding Explicit Information, Integrating Textual Information, Summarizing Textual Information, and Inferring Situational Meaning in addition to two strategies of Keyword Matching and Eliminating Implausible Answers were employed by all the high-proficiency students to answer this item. Test Taking Experience (TTE) was also reported by one high-proficiency student. A relatively similar skill range was reported by at least one student in the mid-proficiency group, while low-proficiency student relied entirely on retrieving literal meaning from a sentence and making inference from that

detail. All students got this item incorrectly. What seems to be missing from all these students' verbal protocols as compared with expert judgment is the use of Understanding Cohesive Devices. The correct answer to this item starts with a discourse marker "on top of that" which manifests itself to be an important linguistic clue that readers should attend to. It is possible that the discourse marker itself is new to the students, and a lack of its knowledge prevents them from making correct interpretation. This is revealed via two students' reports:

Student 2: *"I think 'on top of that' means in conclusion. However, the clause that follows introduce new information. So, I eliminate this option."*

Student 1: *"Question 20, the most appropriate statement to complete the passage. I'm not sure because I don't have sufficient information to answer. I use the elicitation method. I think A and C may be correct. Option A, 'on top of that', it introduces new requirements, not concluding the passage. So, I chose C. It should be the concluding sentence of the passage. I use my previous experience of test taking, at the end, the author usually give advice."*

It can also be inferred from these students' protocols that they might have misinterpreted the question. Rather than understanding "the best sentence to complete the reading passage" as an open question the answer to which may be either a conclusion or a new piece of information ensuing from the passage, they approached the text and the options with a rather restricted perspective – to search for evidence of the conclusion.

Item 24 requires students to identify the main idea of a paragraph. Students' deployment of Understanding Explicit Information, Integrating Textual Information, Inferring Situational Meaning, and Summarizing Textual Information seems to be in line with what was thought by experts. Except for two low-proficiency students, all other students answered this item correctly. Contrary to Item 20 which requires readers to build an inter-textual model of comprehension, Item 24 can be tackled by integrating information within a paragraph, which makes it easier for the students.

In short, out of the three items judged to mainly assess Summarizing Textual Information, Item 20 seems to be the most challenging as all nine protocollers chose the wrong answer. While they have employed the majority of subskills expected by the experts, they missed Understanding Cohesive Devices subskill which plays an important part in connecting the information in the question with that in the text. Item 11 was intended to test readers' ability to integrate and summarise information across sentences. However, some students could answer this item correctly just by recognizing the discourse markers and relate them with the

key words in the options. The skill use pattern to answer Item 24 seems to be consistent across proficiency levels and aligned with expert judgment.

5.3.8. Items assessing Identifying Text Structure

Only one item, which asks about the organization of the reading text, was judged by experts to mainly assess Identifying Text Structure (ITS). Understanding Explicit Information (UEI), Integrating Textual Information (ITI), Summarizing Textual Information (STI), Inferring Situational Meaning (ISM), and Identifying Text Structure (STI) were thought to be involved in the reading process.

High-proficiency students' reading process seems to be consistent with expert judgment as they all employed the subskills identified by experts. An additional strategy of Eliminating Implausible Answers (EIA) was also exercised by one of them. Students in the mid- and low-proficiency groups showed less flexibility in skill use. Only one mid-proficiency student reported using Identifying Text Structure, Integrating Textual Information and Inferring Situational Meaning. Others used a single subskill to answer the item. Only three students, two high-proficiency and one mid-proficiency, identified the correct answer. Others either made an incorrect inference (one high-proficiency and two mid-proficiency) or made uninformed guesses (low-proficiency students). As revealed via their verbal reports, most of the students who answered this item incorrectly were under time pressure and hence could not synthesize information across paragraphs to derive the correct answer.

5.4. Discussion

As an essential component of Chapelle et al. (2008) three-plane model of explanation inference, the exploration of expert judgment and students' reported use of reading skills and strategies in responding to the test items contributes to the understanding of the construct of the test under investigation. Misalignment, if any, between expert judgment of reading skills/strategies and students' actual use of those skills/strategies suggests potential construct irrelevance elements. In contrast, correspondence between expected and actual skill use offers evidence in favour of the meaningfulness of the interpretation of the test score as the item measures what it is purported to measure.

Analysis of expert judgment and students' verbal protocols revealed that there was a large agreement between expert judgement and data from student verbal reports in terms of the primary reading skills assessed by the test items. Out of 40 test items, only four items showed a stark disparity, including one item mainly assessing Inferring Situational Meaning (ISM) and three items mainly assessing Understanding Pragmatic Meaning (UPM). These reading

subskills were not clearly identifiable from students' verbal protocols although they were judged by experts to be primarily tested by the items. In case of the Inferring situational meaning subskill, students answered the item by matching keywords in the question stem with those in the options without much reference to the text. All nine students answered the item incorrectly. Although the disparity does not automatically invalidate the interpretation of the score meaning, it offers *prima facie* evidence of the inappropriateness of the item construction. A closer look at the item content suggests that the item induced the readers to rely more on the interpretation of the options and keyword matching to choose the answer than on making inferences from the text. Specifically, the question asks readers to choose the option that best paraphrases a given sentence in the text. The option that attracted all nine students seemed to be an effective paraphrase of the given sentence, while the correct option contained information that was not stated in the given sentence, but in the previous sentences in the same text. Therefore, via students' protocols, this item manifests itself to be a potential candidate for revision. Three items that were judged to mainly assess Understanding Pragmatic Meaning seemed not to do so: little evidence was observable from students' reports that hinted at the use of Understanding Pragmatic Meaning. Instead, students made inferences based on their understanding of details and information garnered across sentences. This finding, however, does not represent construct irrelevant factors as it reflects the view that pragmatic inference questions are difficult and should not engage readers in making interpretation and inference beyond the textual level of reading (Khalifa & Weir, 2009).

Contrary to the general agreement on the primary reading subskills, disparity becomes more pronounced when it comes to the potential involvement of the subskills in responding to a particular item. The degree of disparity increases as readers proceed from lower level to higher level processes of reading and with respect to different proficiency levels. Items that mainly require readers' inferencing skills – a higher level process of reading, such as Inferring Situational Meaning (ISM), Understanding Pragmatic Meaning (UPM), and Summarizing Textual Information (STI) induced readers to use a diverse range of reading subskills and strategies to deduce the answers. In a similar vein, readers across different proficiency levels, particularly those at two ends of the proficiency spectrum, differed greatly with respect to the skill use patterns. High-proficiency students seemed to employ a range of skills more compatible with what was expected by experts while low-proficiency students primarily drew on their understanding of single sentences. These discrepancies could be explained by both theoretical perspectives and empirical evidence as revealed via the study's findings which are discussed below.

Cognitive processes of reading

Understanding Explicit Information at local level – a lower-level process of reading – was judged by experts to be involved in answering all 40 items since readers need to retrieve meaning from individual sentences before integrating information across sentences or making inferences. However, this subskill was only consistently identifiable from students' protocols when they processed items that mainly require lower-level subskills. Moreover, two lower-level subskills of Lexical Access and Syntactic Parsing only became salient via students' reports in response to items that require students to pinpoint specific details in the text or to paraphrase sentences. As regard students' reading proficiency, low-proficiency students rely more on lower-level processes of reading than high-proficiency students. This reliance becomes more pronounced when they have difficulty processing the sentence or accessing their mental lexicon, which in turn, prevents them from processing higher-level skills effectively. Several implications can be drawn from the above findings. First, high-proficiency readers are likely to achieve automaticity in processing lower-level skills of reading, thereby enabling them to operate higher-level processes more effectively. On the other hand, low-proficiency students' inefficiency in lower-level processes restricts their use of higher-level skills, rendering them more susceptible to making wrong inferences and uninformed guessing. Lower-level processes of reading are more identifiable among low-proficiency readers than high-proficiency readers because reading comprehension breakdowns happen more often among them. Second, high-proficiency students' efficiency and automaticity in processing lower-level reading skills make them less attended to during the verbal reports. Therefore, the identification of lower level skills among these students were challenging, which contradicts with expert judgment. Despite the discrepancies between expert judgment and students' verbal protocols, the test seems to have functioned well to discriminate high and low proficiency students in terms of their cognitive processing.

The impact of test format

All 40 items in the test employ a multiple-choice format, with each question followed by four options. This selected response format partly induced students to use additional strategies beyond the core reading skills identified by experts, the most prominent of which are Eliminating Implausible Answers and test-wise strategies.

Eliminating Implausible Answers was employed by the majority of students, particularly those in high- and mid- proficiency groups, to answer items 14, 34, and 38 and occasionally emerged from students' responses to other items. The purposes for which this skill was utilized by students varied according to the nature of the question per se and the level of

their understanding of the text. Items 14, 34, and 38 asks readers to choose a statement that is not mentioned in the text. Due to the requirement of the question, almost all students reported identifying statements that they can find support evidence in the text, so that what is left is the correct choice. Therefore, they needed to actively engage in processing the text rather than resorting to common sense and test-taking experience to eliminate implausible answers. The strategy operated in this way implies that readers proceeded “via the text” rather than “around the text” (Cohen & Upton, 2007), and that they maintained active control over the text rather than using the strategy to “circumvent the need to tap their actual language knowledge” (Cohen & Upton, 2007).

Another purpose for which the strategy was used by the respondents was to compensate for a lack of the text understanding or to find answers quickly based on the clues in the question stem and options instead of going back to the text. One student reported that she eliminated two options that contain words with extreme meaning, such as absolutely and definitely to give her a 50 percent chance of getting the item correct since she could not locate parts of the text that contained the needed information. The strategy used in this way can be seen as a test wise strategy as little engagement with the text was reported.

The strategy was also employed by some high-proficiency students as a way to confirm their chosen answers. This use happened quite often with items that require readers to find a suitable place in the text for a given statement, and lexical inferencing items. As such, after identifying the most plausible answer, these students replaced the chosen answer with different alternatives and reinterpreted the text to see if it made sense. Although this use of the strategy was explicitly stated by the respondents, it does not constitute a core reading process that they employed to derive the answers. It is, instead, considered a peripheral strategy employed to carefully recheck the answers by students, particularly those in the high-proficiency group.

In addition to the use of Eliminating Implausible Answers as a test-wise strategy, in some cases, respondents drew on their experience in test taking and test practice to answer the items. This is evident in one high-proficiency and one mid-proficiency student’s reports in their responses to items 20 and 26 respectively. The former claimed that he drew on his experience in responding to questions of the same type, herein choosing the best concluding sentence to complete the passage, to derive the answer. The latter relied on her reading experience to locate parts of the text that contain the necessary information to the questions. While the former used the strategy as the last resort to identify the answer, the latter actively employed the strategy to quickly find the best option. The use of these test-wise strategies seems to derail the reading process intended by the test constructors and identified through expert judgment.

Construct irrelevant factors

Some items were judged to mainly assess one skill but elicited other skills as evidenced via students' reports. Of most relevance are lexical inferencing items where respondents are expected to engage in processing the text and use textual cues to infer the meaning of the words. However, in many cases students, particularly those at high-proficiency level, could choose the correct answers without recourse to the text, but instead using their vocabulary knowledge. Although vocabulary knowledge was identified as a critical predictor of lexical inferencing success (Nassaji, 2006), and the ability to make correct inferences depends largely on an adequate knowledge base of vocabulary (Laufer, 1992, 1996; Nation, 1993), that the students could answer the items correctly without attending to the text suggests that the items have failed to elicit the intended reading processes among high-proficiency students. This could have ensued from the inappropriate selection of vocabulary items to construct the test questions, items that are too easy for them to even look for cues in the text to decipher their meaning. In one of the items, students did make an inference, but based their responses entirely on the meaning of the given word and the four options rather than on the textual cues. The process to derive the meaning of this item was, therefore, irrelevant to the one intended by the test designer, thus representing it as a potential candidate for revision.

Understanding Pragmatic Meaning items, though believed by experts to mainly assess readers' ability to infer the attitude, purposes, and tone of the authors, have resulted in students primarily inferring situational meaning from the text. In most of students' reports they made inferences about attitude, tone, and purpose of the author based on clues in the text rather than on their own pragmatic knowledge. Despite this discrepancy between expert judgment and students' reports, the finding seems reasonable given that this is a test of reading comprehension rather than a test of pragmatic knowledge, and that pragmatic inference questions should be designed so that students can rely more on cues in the text to find the answer than on using their pragmatic knowledge. The discrepancies might have resulted from experts' selection and interpretation of the technical terms that informed the judgment process. Accordingly, the use of the term pragmatic meaning and the wording of the questions might have predisposed experts to prioritize Understanding Pragmatic Meaning as the core subskill tested by the items while Inferring Situational Meaning was regarded as an essential subskill in the reading process. Understanding Rhetorical Information might be a more suitable terms to describe the subskill under investigation.

In case of Item 4, as discussed in Section 5.3.6, the wording of the question stem has induced all respondents to use the peripheral subskill of keyword matching as opposed to the

core skill of Inferring Situational Meaning to derive the answer. As a result, all participants answered the item incorrectly. Similar findings were also reported in Jang (2009) who found that readers approaching items that mainly require the use of inferencing skills via basic textual comprehension skills such as keyword matching tended to fail to infer the underlying intention or meaning of a phrase or sentence in the question. In this study, little evidence of textual engagement was recorded from students' verbal reports, making it a potential candidate for item revision.

Local item dependence

Local item dependence in a reading test refers to the mutual dependency among items that share the same reading passage so that response to one item is affected by or dependent on other items (Fan & Bond, 2019). Local item dependence is another threat to the interpretation and use of the test scores, and therefore, should be avoided in any language test. Evidence of local item dependence in the current study emerged from students' reports in response to Items 21, 22, 24, and 25. Item 21 asks about the purpose of a detail mentioned in the text, the answer to which could be used to infer the answer to Item 22 which also asks about the purpose of the same detail. Item 24 is about the main idea of a paragraph. Information garnered to answer this item could also be used to answer the next question – Item 25 which mainly assesses readers' ability to identify the antecedent of a referent word. One mid-proficiency student claimed that she could recognize the coreference between Item 21 and Item 22 because they ask about the same detail. Since the author was describing an unusual experience of the pilot (answer to Item 24), the student could infer that the purpose of the author was to introduce the idea of global warming (answer to Item 25). In a similar vein, another mid-proficiency student reported that she could identify the answer to Item 25 without going back to the text because the information to answer Item 25 has been used to answer Item 24. The main idea of the paragraph asked in Item 24 was about the politicians who took no action against global warming while the antecedent of the referent word in Item 25 was the politicians identified in Item 24. The major source of the disparity between expert judgment and students' protocols, therefore, can be traced back to the practice of item construction where test designers used the same prompt to develop different items. Thus, these items can be considered potential candidates for revision.

In summary this chapter provides a detailed analysis and discussion of the findings from the expert judgment and stimulated verbal recall data to gain insights into the reading processes of the test takers while they took the test and the extent to which they align with what were reported by the experts. The findings offer some useful implications for the teaching and assessing of L2 reading skills, which is discussed in depth in Chapter X.

CHAPTER VI: FACTOR STRUCTURE AND FACTORIAL INVARIANCE OF THE TEST

6.1. Introduction

This chapter reports findings of the second phase of the research which focuses on identifying the underlying factor structure of the L-VSTEP reading test and testing the invariance of this structure across different subsamples. The findings of this phase contribute an additional layer of evidence to the validity of the score interpretation and use of the L-VSTEP reading test by addressing the second plane in the three-plane explanation framework by Chapelle et al. (2008). Of importance here is the examination of the alignment between the theoretically proposed and empirically derived model of the test constructs. Where such alignment is not supported by empirical data, the study offers potential evidence of threats to the construct validity of the test, thereby rebutting the interpretive argument as presented in chapter IV. In addition, once the underlying pattern of the test scores reproduces the theoretically proposed model of the test construct, a more restrictive level of model scrutiny is undertaken. That is to ascertain if the confirmed factor structure remains invariant across different sub-samples of the same data set, herein the high-proficiency and the low-proficiency student groups as indicated by their scores on the test and students across different academic disciplines. If the factor structure of the test is found to be invariant across high-performance and low-performance groups and across groups of students who are expected to use English for different purposes (e.g., English for pedagogy, English for translation, and English for general purposes), the interpretation of the test scores is meaningful regardless of students' L2 reading performance and academic disciplines. Otherwise, caution should be exercised when interpreting and using their scores as they may not represent the same model of the test construct as theoretically defined. As will be unpacked later, analyses of the factor structure and factorial invariance generated both supporting as well as rebutting evidence for the interpretation and use of the test scores. In what follows, a detailed description of the model building process, data analysis, and findings of the study is provided. This is followed by a discussion of the salient findings with reference to relevant literature and implications for the interpretation and use of the test scores.

6.2. Item coding and unit of measurement for Confirmatory Factor Analysis (CFA)

An essential constituent component in the CFA model building and testing process is the identification of reading subskills as specified in the guidelines for test item writers and the relevant test items that are designed to measure those subskills. Since the same reading test (form A) was used in both the stimulated verbal recall phase and the CFA analysis, the expert judgment of the reading subskills and associated test items in chapter V was employed in this phase of the study. Table 6.1 presents the reading subskills, their descriptions, and relevant test items identified in the expert judgment stage, which were later used to inform the specification of CFA models to address the research question in this chapter.

Table 6. 1. Expert judgment of reading subskills and the relevant test items

Reading sub-skills	Descriptions	Items
Understanding Explicit Information at local level (UEI)	The ability to locate and understand explicit meaning at the sentence level.	1, 12, 15, 16, 32
Understanding Cohesive Devices (UCD)	The ability to understand the relationship between sentences or ideas using connective devices such as discourse markers, anaphoric and cataphoric references, substitutions, repetitions.	5, 9, 19, 25, 31
Integrating Textual Information (ITI)	The ability to synthesize information from different parts of a paragraph or a text.	2, 7, 8, 14, 17, 26, 34, 38
Summarizing Textual Information (STI)	The ability to understand main ideas and recognize supporting details at paragraph and discourse level.	11, 20, 24
Inferring Situational Meaning (ISM)	The ability to make inferences about details, relationships, situations, and arguments using textual or background knowledge.	4, 18, 21, 35, 36, 37, 39
Understanding Pragmatic Meaning (UPM)	The ability to understand author's purpose for writing, attitude, tone, mood, belief, and intention of using particular rhetorical techniques.	10, 22, 27, 29, 30

Lexical Inferencing (LI)	The ability to guess the meaning of words using contextual clues.	3, 6, 13, 23, 28, 33
-----------------------------	---	----------------------

Instead of conducting item-level CFA as in previous studies (Kim, 2009; Sawaki et al., 2009), the present study used item parcels as data entry for the analysis of the CFA models. Item parcelling is a technique normally used in structural equation modelling to aggregate individual items into one or more parcels and use these parcels rather than individual items as observed indicators of the latent constructs (Matsunaga, 2008; Song, 2008). Item parcelling has been used in a number of studies delving into the factor structure of language proficiency tests in general (In'nami & Koizumi, 2012; Koizumi & Nakamura, 2016; Sawaki & Sinharay, 2018; Yoo & Manna, 2017) and L2 reading proficiency tests in particular (Song, 2008; Van Steensel et al., 2013). Psychometrically, the use of item parcelling enhances scale communality, reduces random errors, alleviates non-normal distribution of data which is detrimental to common estimation methods such as maximum likelihood, and mitigates the idiosyncratic nuances engendered by individual items (Matsunaga, 2008). From a modelling perspective, item parcelling maximizes a model's construct representation, stabilizes model and parameter estimation, increases model fit, and enhances model parsimony (Matsunaga, 2008; Sawaki & Sinharay, 2018).

In the current study, item parcels were also employed as the unit of measurement in the analysis of the CFA models for several reasons. First, the excessive number of items in the test (40 items) requires more parameters to be estimated, thereby increasing model complexity. This destabilizes the model estimation, particularly when the sample size to parameter ratio is low. By parcelling items, fewer indicators per latent constructs are specified and fewer parameters need to be estimated, resulting in more stable CFA solutions (Bandalos, 2002; MacCallum et al., 1996). Second, all 40 items in the test are dichotomously scored, making the item-level input data binary in nature, and their distribution susceptible to nonnormality (Song, 2008). This violates the assumption of interval-level data and normal distribution for CFA models. Item parcels can be a solution since the aggregated scores of a set of items forming a parcel can be treated as continuous variables (Matsunaga, 2008) in CFA. Yet two limitations associated with item parceling need to be acknowledged. These are the parameter estimation bias in certain circumstances and the misspecification and misrepresentation of the model due to aggregation of items that represent multidimensional latent constructs, the latter of which is

of most concern (Matsunaga, 2008). Therefore, unidimensionality of the items forming a particular parcel is considered a prerequisite for item parcelling (Bandalos, 2002; Little, Cunningham, Shahar, & Widaman, 2002).

In conducting item parcelling technique, the all-item-parcel approach (Matsunaga, 2008) was used in the current study. This approach involves the aggregation of all the items within a scale and the use of this scale score as a sole indicator of a latent construct. The all-item-parcel approach was adopted in this study for both technical and conceptual reasons. Technically, there is an unequal number of items per latent construct (subskill) as suggested by the expert judgment, with one construct having only three item-indicators while other constructs having from five to eight item-indicators. This inflates the variability of the summated scores among parcels. Conceptually, aggregating items into a composite score as a sole indicator of a latent construct is congruent with the design and reporting practice of the L-VSTEP test. Test items are designed to measure particular reading subskills, which was confirmed in the expert judgment and stimulated verbal recall stage. This test design and reporting practice seems to be in line with findings from previous studies on the factor structure of language proficiency tests where a second-order factor of general language ability/skills is explained by several first-order individual skills/subskills (Koizumi & Nakamura, 2016; Sawaki et al., 2009; Song, 2008; Van Steensel et al., 2013). Therefore, the use of composite scores as sole indicators (here the seven reading subskills) of the latent variable (here the overall reading skill) constitutes a more comprehensive representation of the test structure and is in congruence with the literature. More details regarding the specification of the CFA models tested in this study are presented in the next section.

6.3. CFA model building

The study adopted a model competing strategy. This strategy helps preclude the inadvertent exclusion of alternative comparable models that may equally or better represent the underlying test structure. The model building, therefore, starts with the specification of the baseline model which is informed by the test specification and expert judgment. Two competing models are also proposed on the basis of current literature.

Model 1: The general reading proficiency model

This model specifies a general reading proficiency construct explained by seven observed indicators which correspond to the seven subskills identified in the expert judgment stage. Each observed indicator is the parcel score of a reading subskill created by summing students' scores on the relevant items. The model is depicted in Figure 6.1.

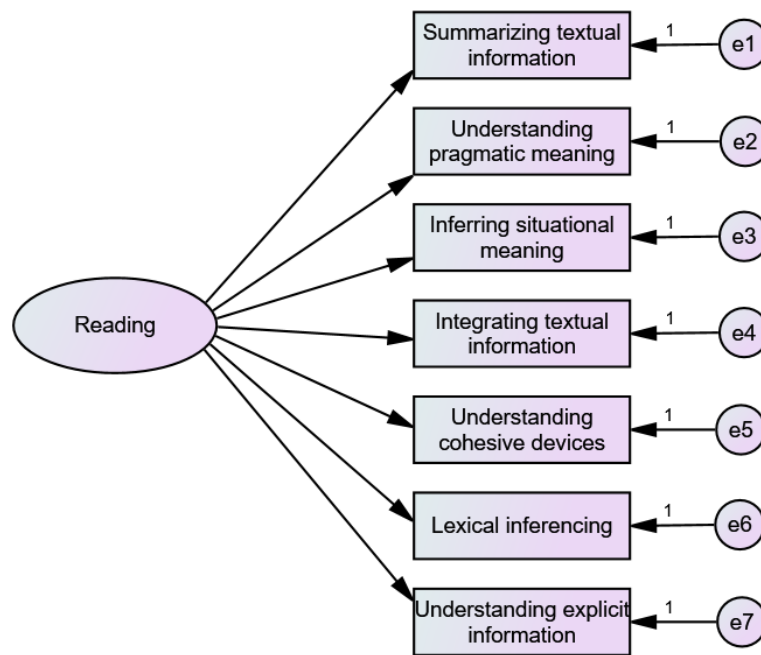


Figure 6. 1. The general reading proficiency model

The latent construct of “Reading” is represented by the circle while the observed indicators are represented by the rectangles. The small circles with a single headed arrow pointing toward the rectangles represent the measurement errors, the amount of variance of each indicator (subskill) not explained by the latent construct of “Reading”. The path loading from the latent construct “Reading” to the observed indicator “Summarising Textual Information” is automatically fixed to the value of 1 in order to resolve the identification problem of the specified model (see chapter IV).

Model 2 and 3: The correlated three factor model and the higher-order factor model

In addition to the baseline model informed by the test specification and expert judgment, two alternative models are specified in Figures 6.2 and 6.3 to explore whether students’ performance on the test may be better accounted for by these models.

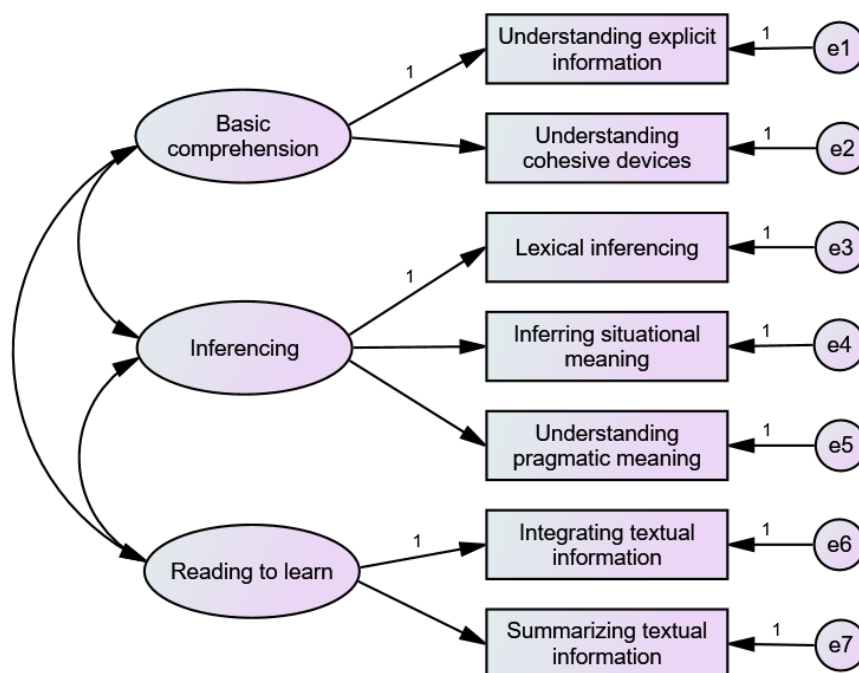


Figure 6. 2. The correlated three factor model

The model in Figure 6.2 depicts three latent constructs which are correlated with each other, namely basic comprehension, inferencing, and reading to learn. The specification of this model was based on the conceptualization of L2 reading construct as reading purposes that informed the development of the reading task specifications for the new TOEFL reading test reported in Cohen and Upton (2006). Therefore, the evaluation of this model yielded potential evidence that enabled the comparison between the underlying structure of the test investigated in the current study and that of the new TOEFL test which can be considered to share similar reading construct – university students’ English reading proficiency. Basic comprehension refers to students’ ability to understand explicit information at the sentence level, factual information, and reference (Cohen & Upton, 2006). This construct is hypothesized to be captured by two lower-level subskills, Understanding Explicit Information and Understanding Cohesive Devices. Inferencing concerns the ability to make inference based on textual clues and to understand rhetorical purposes information (Cohen & Upton, 2006). Lexical Inferencing, Inferring Situational Meaning, and Understanding Pragmatic Meaning are hypothesized as indicators of this construct. Reading to learn involves the ability to recognize the synthesis, categorization and organization of information as well as to understand rhetorical functions such as cause-effect relationship, compare-contrast relationship or arguments in a text (Cohen & Upton, 2006). This construct is hypothesized to be composed of the subskills of Integrating Textual Information and Summarizing Textual Information.

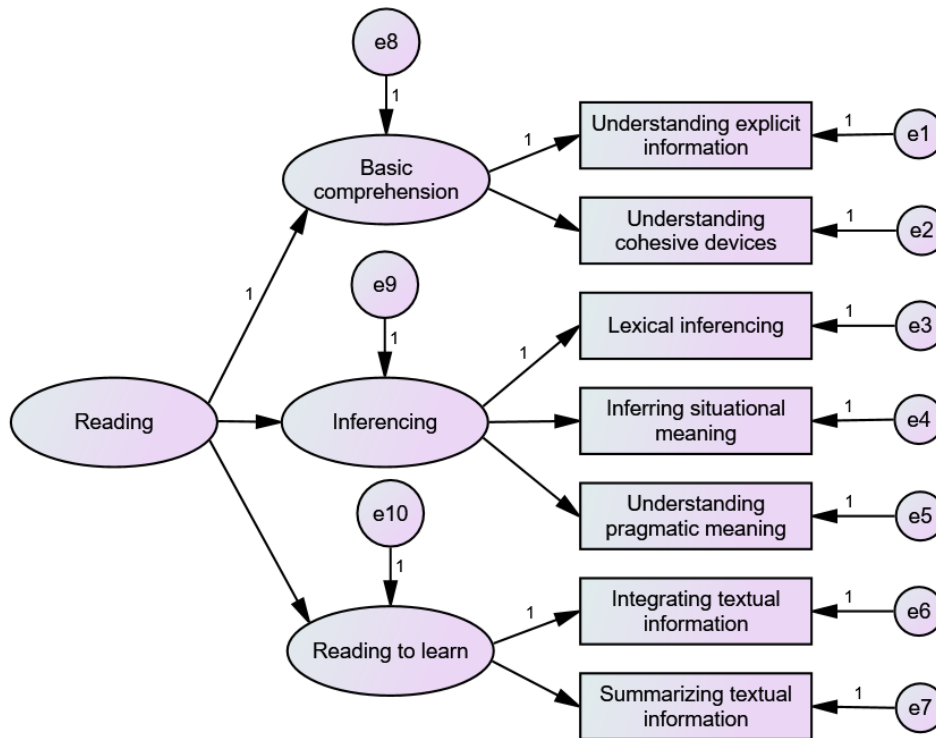


Figure 6. 3. The higher-order factor model

The model in Figure 6.3 is similar to the one in model 6.2 except that the covariance among latent constructs is now fully explained by the higher-order factor of “Reading”. In other words, the three variables of basic comprehension, inferencing, and reading to learn are not specified to be correlated but distinct subskills of reading as in model 2. Instead, they are now hypothesized to be indicators of the general reading proficiency. The specification of models 2 and 3 is also commensurate with those reported in Sawaki et al. (2009) on the factor structure of the Test of English as a Foreign Language (TOEFL IBT) reading test. Sawaki et al. (2009) found that the single trait model (the unitary model) best represented the factor structure of the reading section, suggesting the unidimensionality of the reading ability assessed by the TOEFL IBT test.

Now that the alternative CFA models have been proposed to capture the underlying structure of the L-VSTEP reading test, the following sections present the findings of the analyses and statistically compare the competing CFA models in order to identify which model that the data fit best and therefore best captures the underlying structure of the L-VSTEP reading test.

6.4. Results

6.4.1. Descriptive statistics

Item-level statistics

Table 6.2 presents both descriptive statistics and Rasch measurement results of 40 test items generated by the IBM SPSS software version 23 and WINSTEP Rasch software version 4.4.8 (Linacre, 2020) respectively. Rasch statistics include item fit, item measures, and point-measure correlations.

Table 6. 2. Descriptive statistics (N = 544)

Descriptive statistics			Rasch statistics			Total scores
Item	Correct answer (%)	Measure	Infit MNSQ	Outfit MNSQ	PT-measures	
1	63	-0.33	0.97	0.96	0.32	345
2	53	0.14	0.95	0.94	0.36	290
3	69	-0.6	1.04	1.03	0.21	378
4	20	1.79	1.15	1.19	0.09	111
5	82	-1.33	0.96	0.9	0.28	445
6	45	0.52	1.07	1.08	0.21	244
7	89	-2.14	1.02	1.17	0.09	486
8	38	0.82	1.07	1.08	0.22	206
9	85	-1.56	0.97	0.86	0.25	462
10	67	-0.48	1.04	1.07	0.2	363
11	40	0.66	0.99	1	0.32	216
12	93	-2.42	1.02	1.09	0.1	505
13	66	-0.45	1.04	1.06	0.21	360
14	64	-0.37	0.97	0.94	0.32	349
15	64	-0.36	0.94	0.93	0.35	350
16	27	1.39	0.99	0.98	0.33	147
17	66	-0.45	0.98	0.97	0.3	359
18	53	0.15	0.97	0.96	0.34	289
19	75	-0.88	0.95	0.86	0.33	406
20	24	1.57	1.14	1.19	0.11	130
21	66	-0.44	0.95	0.93	0.35	359
22	28	1.34	0.91	0.92	0.42	151
23	30	1.26	0.95	0.95	0.37	159
24	63	-0.3	1	1.01	0.28	340
25	42	0.64	0.92	0.91	0.42	230
26	66	-0.45	0.95	0.91	0.34	360
27	55	0.07	1.11	1.13	0.14	299
28	53	0.18	1.08	1.11	0.19	286
29	64	-0.34	1.02	1.01	0.24	349
30	61	-0.19	1.04	1.07	0.22	329
31	88	-1.84	0.97	0.87	0.24	479

32	87	-1.74	0.99	0.95	0.2	473
33	54	0.14	0.92	0.9	0.41	289
34	21	1.74	1	1.03	0.29	115
35	35	0.97	0.99	0.99	0.32	190
36	36	0.91	1.02	1.03	0.28	196
37	65	-0.36	0.96	0.95	0.32	350
38	37	0.87	1.08	1.09	0.2	202
39	39	0.79	0.95	0.94	0.38	212
40	33	1.08	0.97	0.99	0.35	179

In terms of the percentage of correct answers, Item 4 has the lowest value (21%), while Item 12 has the highest value (93%). These values indicate that the majority of students answered Item 4 incorrectly and Item 12 correctly (111 vs 505).

Regarding the Rasch statistics, all items have infit and outfit MNSQ values within the acceptable range of 0.6 and 1.4 logits, suggesting that the test items function well enough for their use. It is important to recall that underfitting or overfitting items are causes of concern for test developers as they distort measurement properties. The point measure correlations of the items are all positive, though with low to moderate magnitudes. This suggests that no item functions in an opposite direction to the underlying latent trait explained by the Rasch model (Fan & Bond, 2019). However, some items, such as item 4, item 7, and item 12 have point measure correlations approaching zero, causing concern regarding a competing secondary dimension. A closer inspection of the standardized residual variance after the extraction of the primary Rasch dimension will be reported momentarily to shed more light on the unidimensionality issue.

In the Rasch model, the person ability and item difficulty measures are calibrated onto the same unidimensional linear interval scale, enabling the direct comparison of their relative standings. This information is exhibited in the Wright map in Figure 6.4.



Figure 6. 4. The Wright map (N = 544)

To the left of the vertical line (the linear measurement scale) were students arranged in the order of their performance on the test. Those located at the upper end of the scale were high scoring students and those located at the lower end were low scoring students. Each “#” represents five students while each “.” represents one to four students. Test items can be found on the right side of the vertical line and were presented in the order of difficulty. More difficult items were located towards the upper end of the scale while less difficult items towards the lower end. On both sides of the vertical line, M is the mean, S is one standard deviation from the mean, and T is two standard deviations from the mean. M on the left side was situated higher than M on the right side of the scale, indicating that the mean of the person ability was higher than the mean of the item difficulty. In other words, the test was relatively easy for this particular cohort of students with a larger cluster of students bundled above the mean item difficulty. The test items cover a wide range of difficulty levels, from -2.42 logits (item 12) to

1.79 logits (item 4). On the other hand, the spread of person ability is relatively narrow with the majority of students positioned around one standard deviation from the mean person ability. Several items, such as item 5, item 7, item 9, item 31, item 32, and item 12, were located towards the end of the measurement scale where no students landed, indicating that these items are too easy to measure the targeted participants. The narrow spread of person ability measures and the identification of too easy items that match no students explain the relatively low person separation index (1.53 logits) as presented in Table 6.3, suggesting that more difficult items are needed to distinguish students at different ability levels. There seems to be a gap in the spread of item difficulty measures around the mean person ability. This is suggestive of a lack of items to assess students at this ability level.

Table 6. 3. Separation, reliability and unidimensionality measures

	Separation	Reliability	Variance explained	Eigenvalue of first contrast
Person	1.53	0.70		
Item	10.03	0.99	23.4%	1.90

As can be seen in Table 6.3, the person separation and person reliability indexes are low, augmenting the information in the Wright map that there is a lack of items differentiating students at different ability levels. High item reliability and item separation indexes indicate that the sample size is large enough to reproduce the item difficulty hierarchy, and that the items are widely spread on the measurement continuum. Only 23.4% of the total variance was explained by the Rasch model. According to Linacre (2020) the variance explained by Rasch depends on the spread of the item and person measures. Wider spread of person ability and item difficulty results in a larger amount of explained variance and vice versa. The person separation index of 1.53 logits in the study means that less than two distinct groups of students in the sample were measurable via the test. This is conducive to the low variance accounted for by the Rasch model. Of more importance, however, is the eigenvalue of the first contrast of the residuals. Low eigenvalues (lower than 2 logits) suggest that the possibility of a secondary dimension above and beyond the primary Rasch dimension is negligible (Linacre, 2020). The eigenvalue of 1.90 augmented by the acceptable fit statistics and point-measure correlations indicate that the test is essentially unidimensional, that is the test items measure only one construct of English reading proficiency.

Parcel level statistics

A prerequisite to item parcelling is the unidimensionality of the items forming a particular parcel. This is examined by subjecting each group of items to the Rasch model for dimensionality evaluation. Results are presented in Table 6.4.

Table 6. 4. Dimensionality of the item parcels

Parcel	Item	Infit MNSQ	Outfit MNSQ	PT- measures	Variance explained	Eigenvalue 1st contrast
Understanding explicit information	1	0.95	0.93	0.55	43.5%	1.37
	12	1.03	1.16	0.28		
	15	0.92	0.92	0.56		
	16	1.06	1.28	0.53		
	32	1.01	1.18	0.35		
Lexical inferencing	3	1.03	1.08	0.41	23.1%	1.36
	6	0.99	1.01	0.46		
	13	1.00	0.98	0.44		
	23	0.93	0.94	0.50		
	28	1.07	1.10	0.40		
Understanding cohesive devices	33	0.96	0.95	0.49	35.2%	1.35
	5	0.94	0.90	0.50		
	9	0.97	1.06	0.44		
	19	1.00	0.97	0.53		
	25	1.03	1.29	0.67		
Integrating textual information	31	1.03	1.02	0.38	31.5%	1.30
	2	0.95	0.91	0.48		
	7	1.04	0.90	0.23		
	8	1.04	1.07	0.42		
	14	0.96	0.88	0.45		
	17	0.98	1.08	0.41		
	26	0.98	0.93	0.42		
34	1.04	1.04	0.40			
Inferring situational meaning	38	1.05	1.05	0.41	27.8%	1.38
	4	1.20	1.34	0.27		
	18	0.96	0.96	0.49		
	21	0.97	0.95	0.47		

	35	0.96	0.94	0.49		
	36	1.03	1.09	0.43		
	37	0.99	0.97	0.46		
	39	0.91	0.89	0.53		
	10	1.07	1.11	0.43	25.3%	1.35
Understanding	22	1.03	1.06	0.47		
pragmatic	27	1.02	1.03	0.48		
meaning	29	0.95	0.95	0.52		
	30	0.92	0.90	0.54		
Summarizing	11	0.95	0.95	0.63	30.7%	1.45
textual	20	1.03	1.06	0.54		
information	24	1.02	1.03	0.65		

As can be seen in Table 6.6, all items within each parcel show acceptable infit and outfit MNSQ values in the range of 0.6 – 1.4 logits. As a result, no items were dropped due to erratic score patterns. The point-measure correlations of the items ranged from 0.27 to 0.67 with no negative values. The relatively moderate point-measure correlations suggested that the items within each parcel essentially targeted a similar latent construct. The principal component analysis of Rasch residuals within each parcel yielded relative low eigenvalues of the first contrasts, ranging from 1.30 to 1.45, all lower than 2 logits. These values suggested that the residuals were distributed randomly rather than systematically and that no substantive patterns of residuals existed beyond the primary Rasch dimension. The adequate item fits and point-measure correlations as well as the low eigenvalues within the first contrasts of each measure suggested the unidimensionality of the item parcels. That is to say items in each parcel measure only one underlying construct. This enabled the aggregation of the item scores to represent the parcels in CFA. Preliminary analysis of the parcel scores yielded the statistics presented in Table 6.5.

Table 6. 5. Descriptive statistics of the parcels (N = 544)

Parcels	Mean	Standard deviation	Skewness	Kurtosis
Understanding explicit information	3.35	0.96	-0.218	-0.366
Lexical inferencing	3.15	1.30	-0.053	-0.370
Understanding cohesive devices	3.72	1.04	-0.531	-0.229

Integrating textual information	4.35	1.50	0.177	-0.519
Inferring situational meaning	3.14	1.50	0.234	-0.546
Understanding pragmatic meaning	2.74	1.17	-0.192	-0.425
Summarizing textual information	1.26	0.84	0.076	-0.702
Mardia's coefficient		-2.634		

Both univariate normality and multivariate normality should be examined as essential assumptions prior to running the CFA analyses. The former is a prerequisite to the latter. According to Byrne (2016), skewness has a tendency to influence tests of mean while kurtosis impacts tests of variance and covariance (p.122). Since CFA is based on variance-covariance input matrix, the scrutiny of kurtosis values is of more relevance to the purpose of the study. As shown in Table 6.5, all kurtosis values are lower than the suggested value of 7 (West, Finch, & Curran, 1995) beyond which early signs of departure from univariate normality are evident. The Mardia's normalized estimate of multivariate kurtosis can be considered an indicator of the multivariate normality of the data. Mardia's coefficients higher than 5 indicate that data approximate non-normal distribution (Bentler, 2008). The Mardia's value of -2.63 in this study is suggestive of the multivariate normal distribution of data. This allows for the use of Maximum Likelihood as the estimation method in the analysis of the CFA models.

6.4.2. CFA findings

The evaluation of the hypothesized CFA models was based on four criteria: the appropriate global model fit, adequate parameter estimates, model parsimony, and substantive interpretability of the solutions. Table 6.6 summarises the global model fit indices pertaining to the three hypothesized models of the test's factor structure (see Figures 6.1, 6.2, & 6.3 presented earlier).

Table 6. 6. The global model fit indices

Fit indices	One-factor model (Model 1)	Three correlated factor Model (model 2)	Higher-order factor Model (model 3)
χ^2	11.303	10.378	10.475
p	.662	.497	.574
χ^2/DF	.807	.943	.873
CFI	1.000	1.000	1.000
TLI	1.000	1.000	1.000
SRMR	.019	.019	.019

AIC	39.303	44.378	42.475
BIC	99.488	117.460	111.259
RMSEA	.000	.000	.000
RMSEA confidence intervals	.000 - .034	.000 - .043	.000 - .039

*CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean-Square Residual; RMSEA = Root Mean Square Error of Approximation

*Good model fit is indicated by non-significant χ^2 , normed χ^2 (χ^2/DF) < 3, CFI and TLI > 0.95, SRMR < 0.07, SRMEA < 0.08, and narrow RMSEA confidence interval

The one-factor model

The one-factor model depicts a general latent construct of reading accounting for the variance in all seven observed indicators which were the aggregated scores of the items within each parcel (subskill). This model yielded exceptionally good model fit indices ($\chi^2/DF = 0.807$, $p = 0.662$; SRMR = 0.019; CFI = 1.00; TLI = 1.00; RMSEA = 0.00, CI [0.00; 0.03]). The unstandardized parameter estimates and their associated standard errors, critical ratios and p-values of the one-factor model are presented in Table 6.7, while the standardized estimates are directly exhibited in the model in Figure 6.5.

Table 6. 7. Unstandardized parameter estimates

Weight	Estimate	S.E.	C.R.	<i>p</i>	Error	Estimate	S.E.	C.R.	<i>p</i>
UEI	1.000				e1	0.635	0.048	13.228	***
LI	1.104	0.145	7.618	***	e2	1.338	0.092	14.613	***
UCD	1.120	0.127	8.806	***	e3	0.721	0.056	12.884	***
ITI	1.528	0.178	8.572	***	e4	1.557	0.116	13.387	***
ISM	1.575	0.181	8.703	***	e5	1.526	0.116	13.122	***
UPM	0.761	0.122	6.238	***	e6	1.195	0.077	15.492	***
STI	0.457	0.085	5.371	***	e7	0.643	0.041	15.818	***
					Reading	0.294	0.050	5.865	***

* UEI = understanding explicit information; LI = lexical inferencing; UCD = understanding cohesive devices; ITI = integrating textual information; ISM = inferring situational meaning; UPM = understanding pragmatic meaning; STI = summarizing textual information

The most interesting piece of information in Table 6.7 is the critical ratio (C.R.) which refers to the extent to which the regression weights (factor loadings) and variances (including factor variance and error variances) are significantly different from zero. Critical ratio of a variable is calculated by dividing its unstandardized estimate by its standard error and must be higher than 1.96 to be considered significant at $p < 0.05$. As displayed in Table 6.9, all factor loadings and variances are significantly different from zero at $p < 0.001$ level.

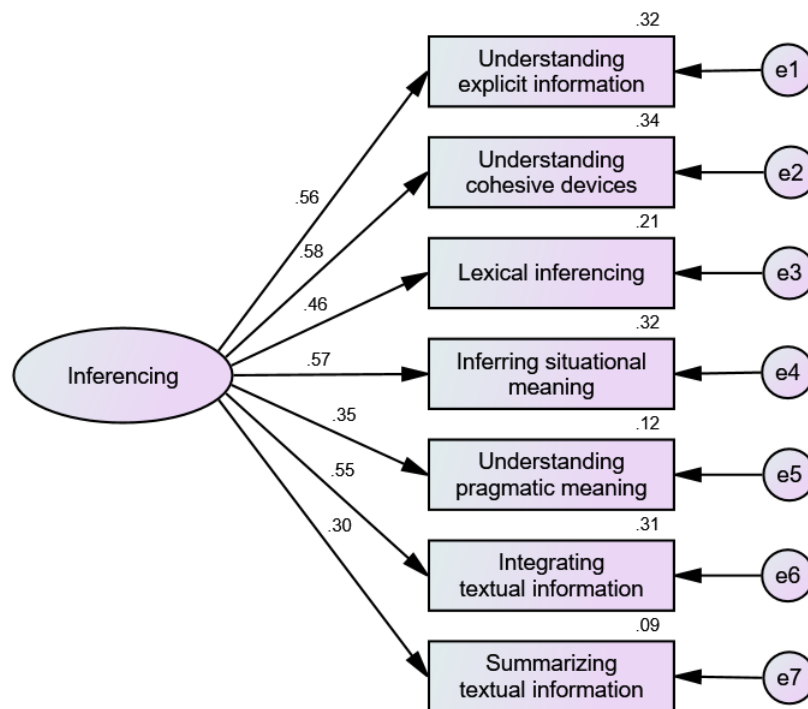


Figure 6.5. The one-factor model with standardized estimates

As shown in Figure 6.5, the standardized factor loadings of the seven indicators are relatively moderate, ranging from 0.30 to 0.58. The latent construct of reading explained a moderate amount of variance in some variables, such as understanding cohesive devices (0.34), understanding explicit information (0.32), inferring situational meaning (0.32), and integrating textual information (0.31) while it accounted for a modest amount of variance in summarizing textual information (0.09) and understanding pragmatic meaning (0.12). These modest explained variances implied that items designed to measure summarizing textual information and understanding pragmatic meaning might have engaged other abilities other than reading proficiency per se.

The three correlated factor model

This model specified three correlated but distinct factors of reading subskills, namely basic comprehension, inferencing, and reading to learn, which may alternatively explain

students' performance on the test. The global model fit indices were reasonably good ($\chi^2/DF = 0.943, p = 0.497$; SRMR = 0.019; CFI = 1.000; TLI = 1.000; RMSEA = 0.000, CI [0.000; 0.043]). The standardized regression weights, variances, and covariances, as presented in Table 6.8 were all significant. However, as can be observed in Figure 6.6, the three correlation coefficients among the latent constructs were excessively high in which there was incidence of a Heywood case ($r > 1$) (Kline, 2016; Schumacker & Lomax, 2010). This indicated that the three latent constructs were too highly correlated to be considered distinct factors. Moreover, since the existence of a Heywood case rendered the solution inadmissible (Kline, 2016), this model was not considered further.

Table 6. 8. The unstandardized estimates of the three correlated factor model

Regression weight	Estimate	S.E.	C.R.	<i>p</i>
UCD \leftarrow BC	1.000			
UEI \leftarrow BC	0.892	0.102	8.721	***
UPM \leftarrow IN	1.000			
ISM \leftarrow IN	2.070	0.332	6.236	***
LI \leftarrow IN	1.460	0.251	5.826	***
STI \leftarrow RTL	1.000			
ITI \leftarrow RTL	3.321	0.627	5.295	***
Variance				
BC	0.390	0.067	5.808	***
IN	0.173	0.049	3.510	***
RTL	0.059	0.022	2.640	**
e1	0.699	0.061	11.432	***
e2	0.618	0.051	12.040	***
e3	1.192	0.078	15.355	***
e4	1.515	0.128	11.813	***
e5	1.329	0.094	14.155	***
e6	0.645	0.042	15.450	***
e7	1.589	0.187	8.507	***
Covariance				
BC $\langle \text{--} \rangle$ IN	0.244	0.042	5.811	***
BC $\langle \text{--} \rangle$ RTL	0.149	0.030	5.017	***
IN $\langle \text{--} \rangle$ RTL	0.105	0.024	4.354	***

* UEI = understanding explicit information; LI = lexical inferencing; UCD = understanding cohesive devices; ITI = integrating textual information; ISM = inferring situational meaning; UPM = understanding pragmatic meaning; STI = summarizing textual information; BC = basic comprehension; RTL = reading to learn; IN inferencing

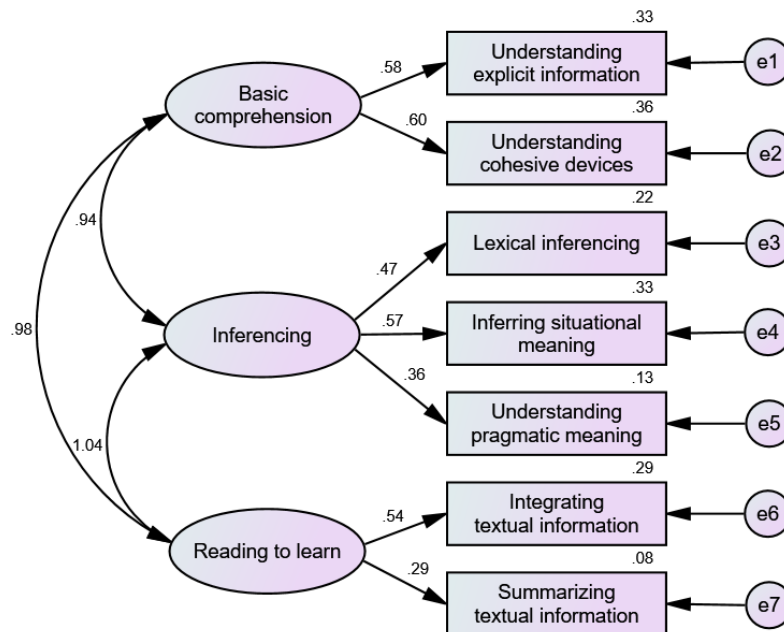


Figure 6. 6. The three correlated factor model with standardized estimates

The second-order factor model

The second-order factor model was proposed as an alternative to the three correlated factor model that generated too excessive correlations among the latent constructs. The three correlations among the latent constructs are now specified to be fully explained by the second-order factor of reading. This model yielded acceptable global fit indices ($\chi^2/DF = 0.873$, $p = 0.574$; SRMR = 0.019; CFI = 1.000; TLI = 1.000; RMSEA = 0.000, CI [0.000; 0.039]).

Table 6. 9. The unstandardized parameter estimates of the second-order factor model

Regression weight	Estimate	S.E.	C.R.	<i>p</i>
BC ← Reading	2.370	0.451	5.258	***
IN ← Reading	1.666	0.371	4.489	***
RTL ← Reading	1.000			
UCD ← BC	1.000			
UEI ← BC	0.891	0.102	8.720	***
UPM ← IN	1.000			
ISM ← IN	2.073	0.333	6.234	***

LI ← IN	1.460	0.251	5.822	***
STI ← RTL	1.000			
ITI ← RTL	3.357	0.628	5.346	***
Variance				
e10	0.000			
e8	0.042	0.049	0.857	0.392
e9	0.000	0.023	0.008	0.993
e1	0.699	0.061	11.432	***
e2	0.618	0.051	12.048	***
e3	1.192	0.078	15.356	***
e4	1.514	0.128	11.801	***
e5	1.329	0.094	14.158	***
e6	0.642	0.041	15.779	***
e7	1.543	0.118	13.024	***

Except for the error variances associated with the basic comprehension and inferencing constructs, all other regression weights and variances were significantly different from zero at $p < 0.001$ level (see Table 6.9). The non-significant error variances of the basic comprehension and inferencing factors implied that the residuals left at the first-order factors as a result of the model fitting process were negligible. In other words, the second-order factor of reading almost perfectly captured the variance in the three first-order factors. This is further elucidated in Figure 6.7 where standardized estimates are presented. Observations among latent factors only (first order and second order factors) indicated that all factor loadings and factor variances were excessively high. The inferencing and reading to learn factors were perfectly captured by the reading factor while factor loadings and factor variances of the basic comprehension construct were approaching the maximum values. These results suggested that all three first-order factors could be merged into the general factor of reading.

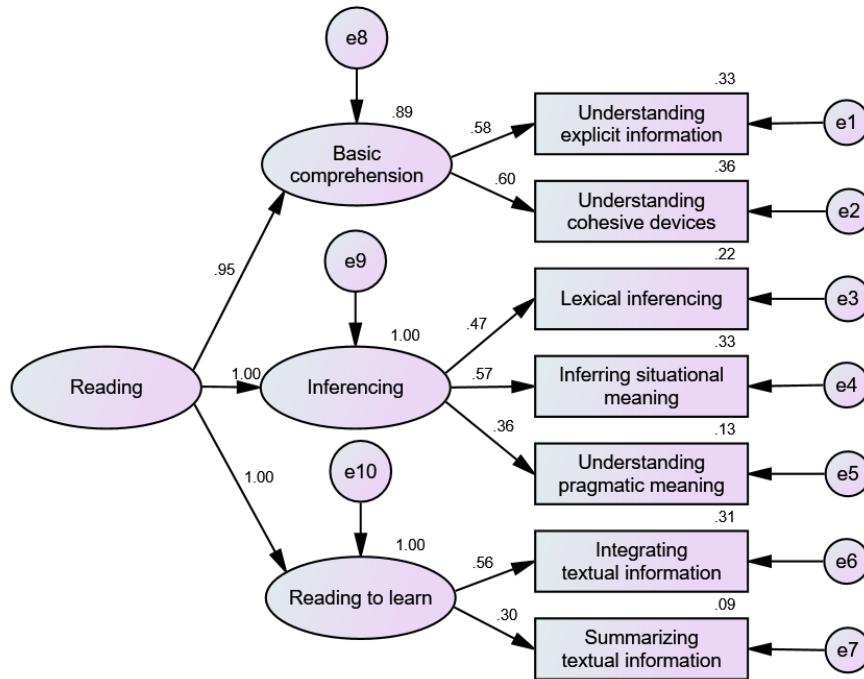


Figure 6. 7. The second-order factor model with standardized estimates

Comparison of the one-factor and second-order factor models

It is recalled that the one-factor model and the second-order factor model have equally good global fit indices. Relatively clear evidence notwithstanding, a structured comparison of the two models would provide an evidence-based ground for the final chosen model. Four criteria for assessing model adequacy as laid out earlier were utilized for the comparison, namely global model fit, parameter estimates, model parsimony, and substantive interpretability.

In terms of global model fit, both models yielded acceptable fit indices. However, the one-factor model shown relative better fit indices than the second-order factor model with narrower RMSEA confidence intervals and lower normed χ^2 value (.807 versus .873). With respect to the parameter estimates, both models generated comparable standardized values. However, two error variances in the first-order factors of the second-order factor model are non-significant, indicating that these factors could be perfectly explained by the second-order factor. Regarding model parsimony, the Akaike’s Information Criterion (AIC) and the Bayes Information Criterion (BIC) were lower in the one-factor model than in the second-order factor model. Both indexes take into account the statistical goodness-of-fit and the number of estimated parameters to penalize less parsimonious models. Finally, the substantive interpretability of the model as a whole and the parameter estimates alluded to the superiority of the one-factor model over the second-order factor model. This is because the loadings and

variances of the first-order factors in the second-order factor model were all approaching maximum values, suggesting that these factors could be merged into the general factor of reading.

6.4.3. Factorial invariance of the one-factor model

The one-factor model which best represented the factor structure of the test was subsequently subjected to a multigroup analysis. The purpose of this analysis was to ascertain if this factor structure was invariant across different sub-samples at the configural and metric levels. Configural invariance refers to the consistency of the model factors and their loading patterns across different sub-samples. In testing the configural model, except for one regression weight which is constrained to one for the purpose of model identification, no equality constraints are imposed on any other parameters in the one-factor model across the sub-groups (Byrne, 2016). Once the configural invariance is established, the analysis proceeds to the metric level where equivalence of the item loadings across sub-groups is examined by imposing equality constraints on all item loadings in the model. The finding of metric invariance indicates that each observed indicator contributes similarly to the latent construct across different sub-groups (Putnick & Bornstein, 2016). On the other hand, metric non-invariance suggests that at least one observed indicator does not function equivalently across the sub-groups. A corollary of this non-equivalence is the testing of the invariance of each item loading separately. This is accomplished by imposing constraints on each individual item loading across sub-groups one at a time and progressively until non-equivalent item loadings are spotted (Byrne, 2016).

Two sets of criteria were adopted to assess factorial invariance. At the configural level, the same set of global model fit indices used to assess the CFA models in the previous stage was adopted. These include the χ^2 statistics, the normed χ^2 (χ^2/df), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR). At the metric level, the significance of the change in the χ^2 value of the two nested models, the configural and the metric models, was used (Byrne, Shavelson, & Muthén, 1989; Reise, Widaman, & Pugh, 1993). A non-significant χ^2 difference value suggests that the metric model is invariant. However, since the χ^2 value is sensitive to large sample sizes, the difference in the Comparative Fit Index was also employed to assess the invariance of the metric model. Cheung and Rensvold (2002) suggested that the CFI difference value should be lower than 0.01 in order for the metric

invariance to be supported. Where the χ^2 difference and the CFI difference values are inconsistent, the latter is used to determine invariance (Gu, 2014).

Two multigroup analyses were performed. The first involves the classification of the students into two groups, a high-scoring group and a low-scoring group. The second was operationalized as three groups depending on the purposes for which English is used, an English for teaching purposes group (hereafter, the pedagogy group), an English for interpretation and translation group (hereafter the translation group), and an English for general communication group (hereafter, the non-English major group). Descriptive statistics pertaining to each sub-group are presented in Table 6.10.

Table 6. 10. Descriptive statistics of the sub-groups

Student groups	N	Percent	M	SD	Minimum	Maximum
<i>English reading proficiency</i>						
High-scoring group	287	52.8	18.20	2.31	11	21
Low-scoring group	257	47.2	26.32	3.75	22	39
<i>Academic disciplines</i>						
Pedagogy group	143	26.3	23.70	6.34	11	39
Translation group	212	39	22.87	4.78	12	35
Non-English major group	189	34.7	19.84	3.33	11	29

In terms of English reading proficiency, the median score of their test paper A result was used to classify student into the high scoring and low scoring groups. The former had a mean score of 18.20 (SD = 2.31) and the latter had a mean score of 26.32 (SD = 3.75). The lowest score in the high-scoring group was 22 while the highest score was 39. Those numbers for the low-scoring group were 11 and 21 respectively.

With regard to academic disciplines, three groups were created, 143 students (26.3%) in the pedagogy group, 212 (39%) students in the translation group, and 189 students (34.7%) in the non-English major group. Their mean scores were 23.70 (SD = 6.34), 22.87 (SD = 4.78), and 19.84 (SD = 3.33) respectively. The lowest and highest scores in each group were 11 and 39, 12 and 35, and 11 and 29 respectively.

Factorial invariance with respect to academic disciplines

The one-factor model was fitted to the three academic discipline groups data simultaneously to test for its configural invariance. All parameters were specified to be freely estimated across the three sub-samples. The configural model fitted the multigroup data fairly

well ($\chi^2/DF = 1.234, p = 0.142$; SRMR = 0.034; CFI = 0.976; TLI = 0.964; RMSEA = 0.021, CI [0.000; 0.038]). The results suggested that the test as a whole yielded similar response patterns by students across different academic disciplines. Since the configural invariance of the one-factor model was supported by empirical data, the next step was to test the metric model where all the factor loadings were constrained to be equal. The metric model can now be considered nested under the configural model because the only difference between the two models is the equality constraints imposed on the factor loadings of the former. The metric model fitted the multigroup data fairly well ($\chi^2/DF = 1.511, p = 0.009$; SRMR = 0.06; CFI = 0.932; TLI = 0.921; RMSEA = 0.031, CI [0.016; 0.044]). Of more importance in the assessment of metric invariance, however, is the significance of the χ^2 difference test and the CFI difference values between the two nested models. These results are presented in Table 6.11.

Table 6. 11. The metric models

Models	χ^2	df	$\Delta \chi^2$	Δdf	<i>p</i>	Δp	CFI	ΔCFI
Configural model	51.844	42			0.142		0.976	
Metric model	81.597	54	29.753	12	0.009	0.003	0.932	0.044
Metric model 1	53.324	44	1.480	2	0.158	0.477	0.977	0.001
Metric model 2	54.434	46	2.590	4	0.184	0.629	0.979	0.003
Metric model 3	60.301	48	8.457	6	0.110	0.206	0.970	0.006
Metric model 4	62.044	50	10.200	8	0.118	0.251	0.970	0.006
Metric model 5	78.282	52	26.438	10	0.011	0.003	0.935	0.041
Metric model 6	67.748	52	15.904	10	0.070	0.102	0.961	0.015

* $\Delta \chi^2 = \chi^2$ difference; $\Delta p =$ significance of the χ^2 difference; $\Delta CFI =$ CFI difference; Metric model 1 = equality constraint on L2; Metric model 2 = equality constraints on L2 and L3; Metric model 3 = equality constraints on L2, L3 and L4; Metric model 4 = equality constraints on L2, L3, L4, and L5; Metric model 5 = equality constraints on L2, L3, L4, L5, and L6; Metric model 6 = equality constraints on L2, L3, L4, L5, and L7.

The χ^2 difference of 29.753 with 12 degrees of freedom yielded a significance value of 0.003 which was statistically significant at $p < 0.05$ level. The CFI difference was 0.044, higher than the suggested value of 0.01 (Cheung & Rensvold, 2002). These results suggested that the one-factor model was not invariant at the metric level across different academic discipline groups. This necessitated the inspection of each specific factor loading to locate the non-

invariant ones. To this end, a series of metric models were specified, each of which had one or more factor loadings held constant across three sub-samples in a cumulative manner. Once a factor loading in a previous model was found to be invariant, it was held constant in the specification of the subsequent models. Figure 6.8 and Table 6.11 help clarify this.

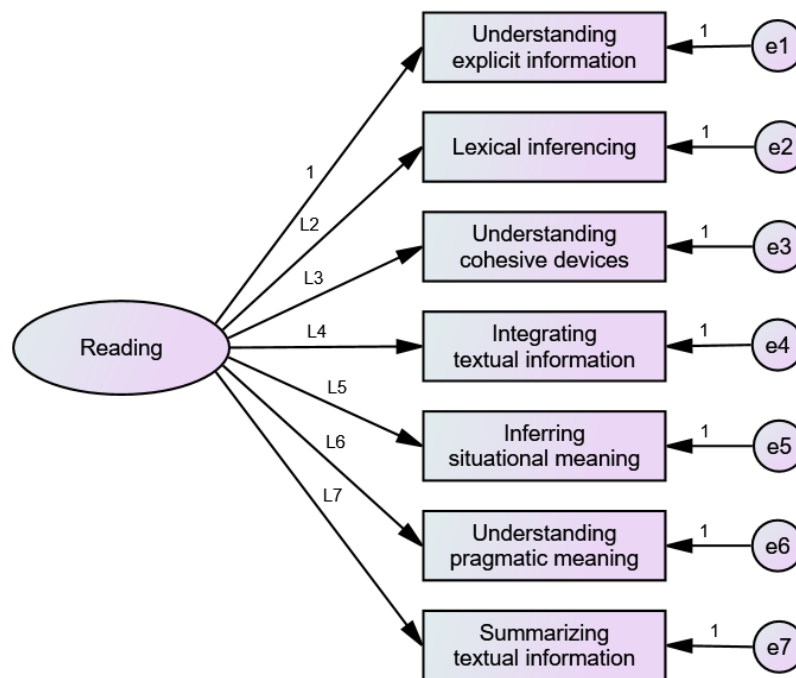


Figure 6. 8. The metric model with equality constraints

In metric model 1, the factor loading of the subskill lexical inference was held constant across three sub-samples. This model yielded a non-significant $\Delta \chi^2$ at $p < 0.05$ level ($p = 0.477$) and a CFI difference value lower than 0.01 ($\Delta CFI = 0.001$). This factor loading was therefore statistically invariant and was held equal in the specification of metric model 2. This procedure was repeated for other metric models. Two non-invariant loadings emerged from this process, the loadings of understanding pragmatic meaning subskill ($\Delta \chi^2 = 26.438$, $p = 0.003$; $\Delta CFI = 0.041$) and summarising textual information subskill ($\Delta \chi^2 = 15.904$, $p = 0.070$; $\Delta CFI = 0.015$) onto the general reading construct. The results suggested that these two sub-skills were construed differently and induced different response patterns by students from different academic disciplines.

Factorial invariance with respect to English reading proficiency

The one-factor model was fitted to the high-scoring and low-scoring groups data simultaneously to test for its configural invariance, following the same procedure as the testing of model invariance regarding academic disciplines. The model did not fit the multigroup data

(χ^2 /DF = 3.203, p = 0.000; SRMR = 0.086; CFI = 0.138; TLI = 0.117; RMSEA = 0.064, CI [0.052; 0.076]). This poor model fit could be attributable to the group-specific nature of the measurement scales (Byrne, 2016, p.254), so much so that the one-factor common model might apply to one group but not the other. This postulation was then tested by fitting the one-factor model to the high-scoring group and low-scoring group separately. Results are presented in Table 6.12.

Table 6. 12. Goodness-of-fit indices for the low-scoring and high-scoring groups

Group	χ^2	p	χ^2 /df	CFI	TLI	SRMR	RMSEA	RMSEA CI
High scoring	8.475	0.86	0.61	1.00	1.00	0.03	0.00	0.00-0.03
Low scoring	57.096	0.00	2.86	0.00	-0.08	0.08	0.09	0.06-0.12

The high scoring group data achieved really good fit. The χ^2 was non-significant at $p < 0.05$ level. The normed χ^2 value was lower than 2. CFI and TLI approached 1 while the SRMR and RMSEA values were significantly lower than the cut-off values of 0.08 and 0.07 respectively. The confidence interval associated with RMSEA value was narrow. These results indicated that the high-scoring group data reproduced the one-factor model as well as the whole data set did. On the other hand, the one-factor model did not fit the low-scoring group data at all. All goodness-of-fit indices significantly deviated from the acceptable values.

In sum, CFA of the test data suggested a one-factor model with reading proficiency as a latent construct measured by the seven subskills identified in the expert judgment stage. This factor structure was tested for its invariance across different sub-samples representing different academic discipline groups and different English reading proficiency groups. The former demonstrated factorial invariance at the configural level but not at the metric level. The latter, on the other hand, retrieved no evidence supporting the invariance of the factor structure of the test. These findings are discussed in light of the relevant literature and L2 reading theories in the following sections.

6.5. Discussion

The purpose of this chapter is to explore the factor structure that best represents the underlying score patterns of the L-VSTEP reading test and challenge the invariance of this structure across different sub-samples in the data. CFA of the competing theoretically proposed models indicated the superiority of the one-factor model that was derived from the test specification over the correlated three-factor model and the higher-order factor model which were informed by the relevant literature. The identification of a one-factor model that most

appropriately represented the test data could be explained from a technical perspective, a language ability perspective, and a psychometric perspective.

Technically, the one-factor model seems to be in line with the technical description and the reporting practice of the test scores. The test is designed to measure test-takers' general English reading proficiency at level B1 to C1 according to the CEFR framework. The specification of the item design suggests seven primary reading subskills which are relevant to the current conceptualization of L2 reading ability in the literature. These subskills were then confirmed in the expert judgment stage, the results of which were reported in chapter 5. The one-factor model seems compatible with this specification in that there is a single underlying factor of reading which is measured by seven observed indicators. Each observed indicator is the parcel score of the items that assess a reading subskill as identified via the test specification and expert judgment. Furthermore, the one-factor model supports the current reporting practice of the L-VSTEP reading test. That is only a single total score of reading is reported with no details of the sub-section (subskill) scores. The finding of a one-factor model of reading and the reporting practice of the L-VSTEP reading test resonated with that of other well-known international English proficiency tests such as the Test of English as A Foreign Language (TOEFL) (Sawaki et al., 2009), the Test of English for International Communication (TOEIC) (In'nami & Koizumi, 2012), the SALT reading test (Van Steensel et al., 2013), and the web-based English as a Second Language Placement Test at UCLA (Song, 2008). It should be noted, however, that Song (2008) and Sawaki et al. (2009) reported reading as a higher-order factor explained by first-order factors of reading subskills while reading in the current study of the L-VSTEP test is specified as a first-order factor measured by observed reading subskill parcel scores. This difference in the model specification is, however, negligible because the higher-order factor model of reading could be considered similar to a one-factor model of reading in that all the reading subskills are subsumed in the underlying higher order reading ability. A favourable outcome of the one-factor model, therefore, also offered support for the higher-order structure of reading ability (Koizumi & Nakamura, 2016).

From a language ability perspective, the factor structure of the reading test identified in the current study neither supports nor declines a clear-cut boundary in terms of the divisibility of L2 reading proficiency. Instead, a third position in favour of a general L2 reading ability with several lower-level reading subskills, which has been established in some previous studies (In'nami & Koizumi, 2012; Sawaki et al., 2009; Song, 2008), seems to gain empirical justification. This factor structure could have ensued from the very test created for the study wherein a general L2 reading ability is the target of assessment while several reading subskills

are specified to inform the item creation process. Similarly, this factor structure provides added value to Alderson (2000) and Song (2008) argument that whether the divisibility of L2 reading ability is assumed or not, it is a common practice among practitioners to consider different reading comprehension subskills or levels of understanding in the building up of reading syllabus or describing readers' competency.

From a psychometric perspective, the acceptance of the one-factor model and the rejection of the correlated three factor model and the higher-order factor model based on consideration of model fit and item parameters implied that reading is a unitary concept, at least to the extent that the current study is concerned. This pattern was also observed in Van Steensel et al. (2013). A possible explanation could be that rather than using a single subskill to answer an item, the test takers might have employed an integrated range of subskills/processes to derive the answer, a phenomenon commonly observed in reading comprehension tests (Rupp et al., 2006). This integrative use of subskills/processes was also emphasized in the cognitive processing perspective of reading. That is reading is a cumulative process with higher-level subskills building on lower-level subskills in an interactive manner. Where comprehension breakdowns occur, competence at one level might be brought to bear on deficiency at the other (Nassaji, 2003). As items in the test were designed to assess a variety of reading subskills, the majority of which were at higher levels of reading, test takers needed to exercise a range of subskills to find the answers. It is fair to assume that in order to answer an inference item correctly, test takers should be able to decode the literal meaning at the sentence level, integrated the decoded information with the incoming information or their prior knowledge, and used other linguistic clues in the text to make inferences. As such, this process involved multiple subskills operated in a systematic sequence. Yet, the extent to which test takers could manoeuvre an integrative skill range depends on how easy the test was and how efficiently they used the reading subskills at their own disposal (Alderson, 2000; Song, 2008; Van Steensel et al., 2013). Since the test was rather easy for the particular sample in the current study as indicated by the Rasch analysis, reading comprehension breakdowns might have happened less frequently, allowing students to integrate different reading skills at different levels smoothly to answer the test items. This might partly explain the integrated skill use pattern as revealed via the CFA analysis.

Another important issue addressed in this chapter is whether the factor structure of the test remains invariant given different sub-samples of the same population. This was achieved via the multigroup analysis where the factor structure of the test was simultaneously subjected to different subgroups of students based on their test performance scores and academic

disciplines. The former addressed the issue of whether the reading test triggered similar response patterns among students of different reading proficiency levels while the latter tested the same hypothesis against students of different academic disciplines.

Results indicated that the one-factor model of reading achieved configural invariance but not metric invariance when students' academic disciplines were taken into account. In other words, students of three different academic disciplines responded to the test similarly at the structural level but not at the item level. A more stringent analysis of the metric models yielded two non-invariant factor loadings, that of the understanding pragmatic meaning subskill and of summarizing textual information subskill onto the reading construct. Both of the subskills are at higher levels of reading comprehension.

Theoretically, pragmatic understanding requires readers to not only process linguistic features of the text but also to be aware of the different pragmatic factors such as written discourse conventions, cultural aspects, text structure, and memory schema (Bensoussan, 1986) between the target language and the readers' mother tongue. As pragmatic understanding is an inferential comprehension process, the possession of an adequate pragmatic knowledge or lack thereof may predispose readers to rely heavily on the textual features (textually explicit), a combination of textual data and pragmatic knowledge (textually implicit), or exclusively on pragmatic knowledge (scriptally implicit) (Alptekin, 2006). The differential responses to the pragmatic items by students across different disciplines in this study might, therefore, be attributed to the differing levels of pragmatic knowledge that they brought to the reading texts. It might be due to the disparity in the English curricula for students of different disciplines where a pragmatic component was featured in one but not in the others. It could also be explained by the fact that the genres and types of written discourse with different cultural specificities that students were exposed to during their English learning process differed across different academic disciplines. From a practical viewpoint, the above-mentioned speculations seem to be warranted given that the pragmatics module was offered to students pursuing bachelor's degrees in English teaching and English translation, but not to non-English major students at the institution where the study was conducted. In addition, as the English curricula were tailored to students who were expected to use English for different purposes, the reading texts, topics, and genres that they were required to read during their study program were likely to be different, resulting in disparity in terms of their pragmatic understanding.

The different text types and genres of English written discourse to which students across different disciplines were commonly exposed might have also borne on the non-invariant factor loading of the summarising textual information subskill. The research literature on reading

comprehension has pointed out that text types and genres did have a role to play in readers' comprehension of main ideas and text structure (Graesser, McNamara, & Louwerse, 2003; Wang, 2009; Yoshida, 2012). Narrative texts, for example, are assumed to trigger the understanding of highly connected ideas which enables the recognition of key propositions and main concepts of the passages while expository texts carry less argument overlaps and more abstract information, thereby blurring the conceptual nodes and inhibiting the extraction of main ideas (Yoshida, 2012). Students of different disciplines, therefore, might have approached textual summarization items differently, and hence responded to the items in a unique way. Empirically supported arguments notwithstanding, the speculative explanations regarding the non-invariant loadings of understanding pragmatic meaning and summarizing textual information subskills onto the general reading construct should be augmented with more empirical findings from replication research and deeper exploration of the multitude of linguistic and pragmatic features of reading texts, which is beyond the scope of this project.

The one-factor model of reading was found to be non-invariant across two groups of English reading proficiency. Further analysis of the data revealed that the one-factor model did not fit the separate samples equally well. While data associated with the high-scoring subsample achieved really good fit, the low-scoring subsample did not support the model, suggesting that the reading test structure held true for the high-proficiency group, but not for the low-proficiency group. A plausible explanation could be that low-scoring students were limited in their range of reading subskills, particularly those at higher levels of reading. They might be either devoid of or inefficient in the use of higher-level subskills, forcing them to over-rely on lower-level subskills in search of the answers to the reading items. This overreliance on lower-level subskills might have either precluded the integrated use of subskills in the comprehension process or driven them to resort to wild guessing strategies when comprehension at lower levels broke down. Neither of the scenarios mentioned above contributes to the finding of a decent relationship among the observed variables in the data, conducing to poor fit of the model. Another potential reason was that the test was too difficult for this particular group of students, making the group more homogeneous and rendering the score range too narrow due to the floor effects (Van Steensel et al., 2013). This low variation in scores in turn reduced the correlation among the variables, thereby making it more difficult to discern the score patterns (Schumacker & Lomax, 2010). Had an easier test been administered to this group, the results could have been different.

CHAPTER VII: TEXT AND ITEM FEATURES AS PREDICTORS OF ITEM DIFFICULTY

7.1. Introduction

This chapter reports findings related to the linguistic and discorsal characteristics of the reading texts and question items as well as the extent to which these characteristics contribute to the item difficulty of the L-VSTEP test. In so doing, the chapter offers validity evidence pertaining to the most concrete plane in the three-plane explanation framework proposed by Chapelle et al. (2008). The exploration of text and item features as predictors of item difficulty reported in this chapter contributes to a better understanding of the substantive meaning of the test construct (Barkaoui, 2015; Gorin & Embretson, 2006). The general assumption is that since the test is designed to assess test-takers' understanding of the reading comprehension passages, the variability in their test scores should be more associated with the linguistic and discorsal features of the texts than with the item features. In case the assumption is not supported by the findings, it is likely that test-takers' performance on the test is more substantially influenced by the linguistic properties of the questions than those of the texts, hence rebutting the interpretive argument as articulated in Chapter IV. The following sections, therefore, provide a descriptive overview of the sample used for this phase of the study and the descriptive statistics pertaining to the linguistic and discorsal features of the reading passages and test questions as well as the item difficulty of the tests. Results of the correlation analyses and multiple regression analyses are then presented to inform the answers to the research question:

What are the linguistic and discourse characteristics of the texts, items and item-text variables of the L-VSTEP reading test? To what extent do they contribute to item difficulty?

7.2. Descriptive statistics

7.2.1. Linguistic and discourse features of reading texts

In chapter IV, linguistic and discorsal features of a test have been thoroughly discussed. Descriptive statistics of the linguistic and discorsal features of the four equivalent test versions of the L-VSTEP test is presented in Table 7.1. Statistics was generated from the text, item, and item-text features of the 16 reading texts (four per test, four tests in total) and

160 items (10 per reading text) used in the current study. Each of these features was used as a variable in subsequent analyses.

Table 7. 1. Text and item features

Variables	Mean	SD	Min	Max
<i>Text length</i>	493.25	29.86	450	545
Syntactic complexity				
<i>Sentence length</i>	20.08	3.37	12.67	24.68
<i>Left embeddedness</i>	4.85	2.00	1.86	9.67
<i>Noun phrase density</i>	0.94	0.14	0.62	1.10
Lexical features				
<i>Lexical diversity</i>	110.07	24.41	74.00	152.45
<i>Word frequency</i>	2.13	0.12	1.95	2.44
<i>Word familiarity</i>	587.43	3.60	580.42	596.06
Referential cohesion				
<i>Content word overlap</i>	0.07	0.03	0.01	0.12
<i>Argument overlap</i>	0.44	0.15	0.13	0.68
Conceptual cohesion				
<i>LSA sentence adjacent</i>	0.21	0.08	0.07	0.35
<i>LSA sentence all</i>	0.18	0.08	0.06	0.37
<i>Connective density</i>	92.89	10.53	75.51	114.00
<i>Text concreteness</i>	379.89	20.52	347.91	422.61
Text readability				
<i>Flesh Reading Ease</i>	50.23	13.58	20.17	74.52
Item features				
<i>Item length</i>	40.70	20.21	12.00	103.00
<i>Item word familiarity</i>	586.96	11.90	538.17	612.06
<i>Item word frequency</i>	9.45	3.65	1.80	22.90
<i>Lexical overlap between distractors and correct answers</i>	0.17	0.13	0.00	0.64
Item-text variables				
<i>Number of plausible distractors</i>				
<i>Lexical overlap between the correct answers and the texts</i>	0.01	0.01	0.00	0.06

Except for the item-text variable of plausible distractors that required expert judgment, all other variables were subjected to the automatic textual analysis Coh-Metrix that generated interval-level data. Text length ranged from 450 words to 545 words with a mean of 493.25. Some variables such as sentence length, the number of words preceding the main verbs, lexical diversity, connective density, text concreteness, item length, and item word frequency showed large variation while other variables such as noun phrase density, word frequency, referential cohesion, and latent semantic analysis of the sentences had small to moderate variation. The number of plausible distractors was the only variable that required expert judgment. However, due to low agreement among the two experts (75.6%) in terms of the total number of distractors in each item, the dichotomous coding scheme proposed by Ozuru et al. (2008) was adopted in this study. Accordingly, items with at least one distractor that could be confirmed/disconfirmed in the texts were coded 1 while items with no such distractors were coded 0. The proportion of items that had plausible distractors and those that did not have any plausible distractors was 2.33 (112/48), with 96.25 percent of agreement among the two experts. Where disagreement occurred (6 items), the experts discussed with each other to reach final agreement.

7.2.2. Item difficulty via Rasch modeling

Students' scores on the test items were calibrated by the dichotomous Rasch model (Bond & Fox, 2015). Table 7.2. presents statistics pertaining to item measures, item fit, item reliability and item separation generated by the Rasch analysis.

Table 7. 2. Descriptive statistics of the item properties

	Mean	SD	Min	Max	Range
<i>Item measures</i>	0	1.20	-4.66	2.70	7.36
<i>Infit MNSQ</i>	1	0.09	0.74	1.34	0.60
<i>Outfit MNSQ</i>	1	0.17	0.50	1.46	0.96
<i>Item reliability</i>	0.96	0.02	0.93	0.99	0.06
<i>Item separation</i>	6.01	2.76	3.74	10.03	6.29

In Rasch analysis, the mean item difficulty is set by default at zero. A negative value of the item difficulty indicates that the test item is easier than the average test difficulty. On the other hand, a positive value suggests that the item is on average more difficult than the overall test. The item difficulty in this study ranged from -4.66 to 2.70, spanning 7.36 logits on the latent trait scale. Infit and outfit MNSQ are both within the acceptable range of 0.5 – 1.5 for

productive measurement and the mean values approach 1, implying that the items functioned according to the expectations of the Rasch model. The mean item reliability was 0.96 suggesting that the item difficulty hierarchy can be confidently replicated if the test is given to a similar test population (Green, 2013). Via the Rasch model, the items were reliably separated into more than 6 levels of difficulty.

7.3. Findings of the correlational and multiple regression analyses

Following Freedle & Kostin’s (1999) proposal for item difficulty modelling of reading comprehension tests, a three-step statistical analysis approach was adopted in the current study. First, the correlations between item difficulty and text, item, and item-text variables were examined. This is to address the criticism levelled at the contribution of the reading text features to item difficulty in reading comprehension tests with multiple choice question format. Unless there is an association between textual features and item difficulty, the level of engagement of the test-takers in processing the reading texts can be questioned. In the second stage, text, item and item-text variables that have a significant correlation with item difficulty were included in multiple regression analyses. Initially, standard multiple regression analysis was conducted to examine the relative contribution of each predictor variable to item difficulty. The expectation was that text and item-text variables accounted for more variance in the item difficulty than did item variables. Finally, hierarchical regression analysis was performed to explore the extent to which text and item-text variables remained significant predictors of item difficulty after partialing out the effects of item variables. Findings related to each of the steps mentioned above are presented in the following sections.

7.3.1. Correlation analysis

Table 7.3. presents results of correlation between the text, item, and item-text variables and item difficulty.

Table 7. 3: Correlation between text, item, and item-text variables and item difficulty

Variables	Item difficulty
<i>Text length</i>	.07
Syntactic complexity	
<i>Sentence length</i>	.09
<i>Left embeddedness</i>	.05
<i>Noun phrase density</i>	-.02
Lexical features	

<i>Lexical diversity</i>	.06
<i>Word frequency</i>	-.01
<i>Word familiarity</i>	-.06
Referential cohesion	
<i>Content word overlap</i>	.01
<i>Argument overlap</i>	.05
Conceptual cohesion	
<i>LSA sentence adjacent</i>	.04
<i>LSA sentence all</i>	.03
<i>Connective density</i>	-.01
<i>Text concreteness</i>	-.17 *
<i>Flesh Reading Ease</i>	-.12
Item features	
<i>Item length</i>	.30 **
<i>Item word familiarity</i>	-.07
<i>Item word frequency</i>	.09
<i>Lexical overlap between distractors and correct answers</i>	.20 *
Item-text variables	
<i>Number of plausible distractors</i>	-.45 **
<i>Lexical overlap between the correct answers and the texts</i>	.09

*. Correlation is significant at the 0.05 level (2 tails)

**. Correlation is significant at the 0.01 level (2 tails)

Out of 21 text, item, and item-text variables examined, only four variables had a significant correlation with item difficulty. These included one text variable of text concreteness, one item-text variable of plausible distractor, and two item variables of item length and lexical overlap between the correct answer and the distractors. The magnitude of the correlation coefficients ranged from low to moderate. The item-text variable of number of plausible distractors had the strongest correlation with item difficulty ($r = -.45, p < 0.01$), followed by item length ($r = .30, p < 0.01$), lexical overlap between correct answers and distractors ($r = .20, p < 0.05$), and text concreteness ($r = -.17, p < 0.05$). The results suggested that items with higher difficulty levels were associated with texts having more abstract items and fewer distractors that can be directly confirmed or disconfirmed in the reading texts, items that had more words, and items with larger overlap between the correct answer and distractors. The correlation analysis only indicated the direction and magnitude of the relationship between

each individual variable and item difficulty. The relative contribution of each of the text, item, and item-text variables as well as the collective contribution of these variables to item difficulty were, however, not modelled. This is the focus of the next section where results of the regression analyses are presented.

7.3.2. Regression analyses

The four text, item, and item-text variables that had a significant correlation with item difficulty were employed as predictor variables in the regression analyses while item difficulty was the outcome variable. As prerequisites to the multiple regression analyses, key statistical assumptions need to be examined. This is to ensure that the estimation of regression models is free from bias and to enhance the generalizability of the findings (Field, 2009). These assumptions include multicollinearity, homoscedasticity, independent errors, normally distributed errors, and the absence of outliers, each of which is discussed below.

Assumptions

Multicollinearity refers to the excessive high correlation among predictor variables. The presence of multicollinearity in a regression model will generate untrustworthy regression coefficients, limit the magnitude of the multiple correlation between the predictors and the outcome, and obscure the individual importance of predictor variables (Field, 2009). Multicollinearity can be detected by examining the correlation among the predictor variables. Correlations higher than .90 is initial evidence of multicollinearity. Variance Inflation Factor (VIF) and the tolerance statistic are also multicollinearity diagnostics. VIF values higher than 10 (Myers, 1990) and tolerance statistic below 0.2 (Menard, 1995) are indicators of multicollinearity. None of the predictor variables in this study had correlation higher than .90 with others. The highest correlation was between item length and lexical overlap between the correct answer and distractors ($r = .642, p < .01$). As can be seen in Table 7.5, all tolerance statistics and VIF values were within the acceptable thresholds.

Independent errors refer to the uncorrelated residual terms between two adjacent observations in the regression model. The Durbin-Watson test can be used to examine the correlation between two adjacent residuals. This test statistic generates values between 0 and 4. Values below 1 and above 3 are indicators of violation of the assumption, while values approaching 2 are desirable (Field, 2009). The Durbin-Watson statistic in the current study was 2.02 which was close to 2 and suggested that the assumption of independent errors was satisfied (see Table 7.4).

The detection of outliers is also important in regression analysis as outliers affect the estimation of regression coefficients and make the model biased (Field, 2009). Outliers can be examined by consulting Cook's distance – the influence of a particular case on the overall model, and the standardized residuals – the difference between the outcome variables as predicted by the model and the outcome variable as observed in the data. Cook's distance values below 1 are considered acceptable (Cook & Weisberg, 1982), while the absolute values of standardized residuals should not exceed 2.58 in more than 1 percent and 1.96 in more than 5 percent of the sample cases (Field, 2009). The maximum value of Cook's distance in this study was 0.07 while one sample case (0.6%) had standardized residual value above $|2.58|$ and 6 cases (3.75%) had standardized residual values above $|1.96|$. These results suggested the absence of outliers that may unduly influence the regression model.

Homoscedasticity, also referred to as homogeneity of variance, is the assumption of the equal variance of the residual terms at each level of the predictor variables. Violation of this assumption may inflate Type I error and lead to false positives. Linearity indicates the linear relationship between the predictor and the outcome variables in a regression model. If this assumption is violated, the generalizability of the model will be limited (Field, 2009). The assumptions of homoscedasticity and linearity can be examined by inspecting the plot of the standardized residuals against standardized predicted values.

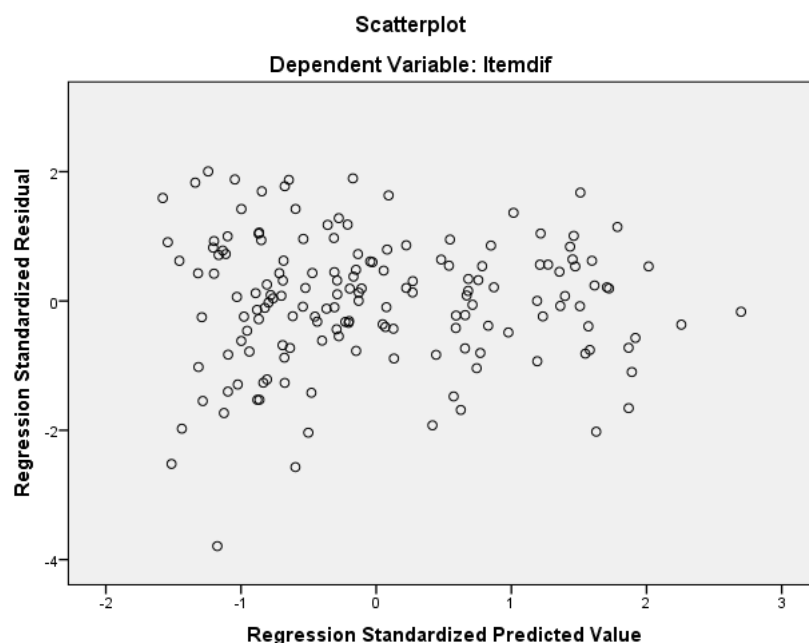


Figure 7. 1. The scatter plot of standardized residuals

As observed in Figure 7.1, the dots appear to be randomly spread out and evenly dispersed around zero. There were no obvious observations of a funnel shape, a curve, or both

in the graph. This is evidence that the assumptions of homoscedasticity and linearity have been met.

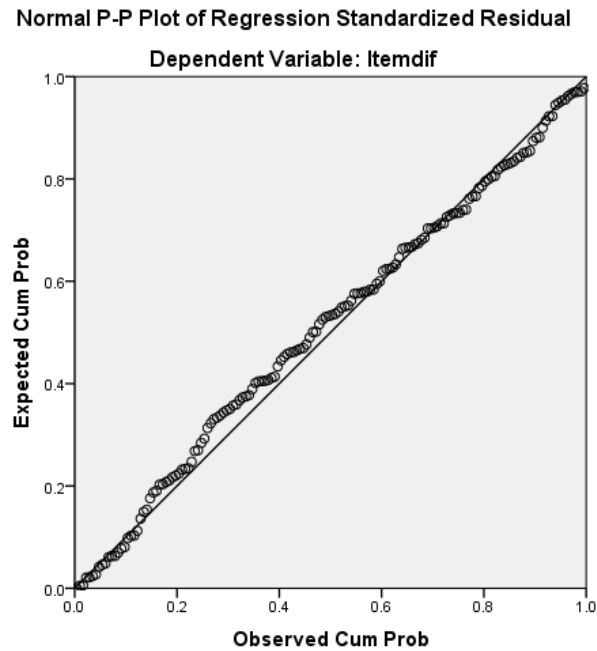


Figure 7. 2. The P-P plot of standardized residuals

The assumption of normally distributed errors means that the errors of the estimation process are minimum and very few errors are greater than zero (Field, 2009). The normal probability plot in Figure 7.2 can be used to examine this assumption. The straight line represents normal distribution while the dots are standardized residuals. As can be observed from the graph, the residuals were positioned closely along the straight line suggesting no obvious deviation from normal distribution.

All in all, the examination of the key statistical assumptions suggested that the input data were appropriate for the regression analyses.

Standard regression analysis

In the first regression model, the standard procedure was adopted. That is all predictor variables were admitted into the regression model simultaneously. The relative contribution of each predictor variable can be considered as the unique variance accounted for by a particular predictor above and beyond the variance accounted for by others. Results of the standard regression analysis are presented in Tables 7.4 and 7.5.

Table 7. 4. Model summary

Model	Model summary				ANOVA		Durbin - Watson
	R	R square	Adjusted R square	Standard error	F	Sig.	

1	0.536	0.287	0.269	0.982	15.519	0.00	2.021
---	-------	-------	-------	-------	--------	------	-------

The model with four predictor variables of plausible distractors, lexical overlap, item length, and text concreteness was statistically significant different from the hypothetical model where no predictor variables were entered, $F(4, 154) = 15.519, p < .001$, and accounted for 28.7 percent of the variance in item difficulty (see R square in Table 7.4). The adjusted R square denotes the variance explained by the predictor variables if the model is generalized to the population. In this study, the model explained 26.9 percent of the variance in the population. Therefore, if the model were obtained from the population rather than the specific sample used in the study, the predictive power of the model would reduce by 1.8 percent.

Table 7.5 displays the parameter estimates of the regression analysis. The unstandardized regression coefficients indicate the extent to which each individual predictor influences the item difficulty, controlling for the effects of other predictors. For example, the B value for the predictor variable of plausible distractor was -1.07, indicating that one unit increase in the plausible distractor resulted in 1.07 unit decrease in the item difficulty given all other predictors were held constant. The regression equation can be rewritten as follow:

$$\text{Item difficulty} = 3.23 + (-1.07 * \text{plausible distractor}) + (0.15 * \text{lexical overlap}) + (0.01 * \text{item length}) + (-0.01 * \text{text length})$$

While the unstandardized regression coefficients represent the weight of each individual predictor variable, their relative contribution to the regression model cannot be compared due to differences in the units of measurement. The standardized regression coefficients were also yielded to supplement the interpretation of the model. These coefficients were standardized so that each had a standard deviation of 1 and a mean of zero, hence allowing for the direct comparison of the predictor variables. As can be seen in Table 7.5, the strongest regression weight was for plausible distractor ($\beta = -0.43$), so that for each standard deviation increase in plausible distractor, the item difficulty reduced by 0.43 standard deviation. The standardized regression weights for lexical overlap, item length, and text concreteness were 0.02, 0.23, and -0.14 respectively. Except for lexical overlap, all other predictors were significantly different from zero. However, caution should be exercised when interpreting the regression coefficient of the text concreteness variable. The 95% confidence interval for the B value of this variable crossed zero (-.016 – .00), suggesting that its regression coefficient might not be reliable (Field, 2009; Jeon, 2015).

Table 7. 5. Parameter estimates

<i>Predictors</i>	<i>Unstandardized</i>		<i>Standardized</i>			<i>Correlations</i>		<i>Collinearity</i>	
	<i>coefficients</i>		<i>coefficients</i>					<i>statistics</i>	
	<i>B</i>	<i>SE</i>	<i>Beta</i>	<i>t</i>	<i>Sig.</i>	<i>Zero-order</i>	<i>Partial</i>	<i>Tolerance</i>	<i>VIF</i>
Constant	3.23	1.49		2.17	0.03				
Plausible distract.	-1.07	0.17	-0.43	-6.26	0.00	-0.45	-0.45	0.99	1.01
Lexical overlap	0.15	0.76	0.02	0.20	0.84	0.18	0.02	0.58	1.73
Item length	0.01	0.01	0.23	2.53	0.01	0.29	0.20	0.58	1.74
Text concrete.	-0.01	0.00	-0.14	-2.05	0.04	-0.18	-0.16	0.97	1.04

Hierarchical regression analysis

In the hierarchical regression analysis, variables were entered into the model by blocks in a sequential manner. The order by which the predictor variables were admitted into the model depended on theoretical assumptions rather than on statistical hypotheses. Given that the chapter was focused on exploring the extent to which the text variables contribute to item difficulty as opposed to item and item-text features and whether the text features remain significant after the effects of item and item-text features are controlled for, it was determined that the item and item-text features be entered in the first block and text features in the second block. Now that the lexical overlap variable was found to be insignificant in the previous model, it was excluded from further analysis. The hierarchical regression model, therefore, included three predictor variables, the item and item-text variables of item length and plausible distractors were entered first, followed by the text-related variable of text concreteness. Results of the hierarchical regression analysis are exhibited in Table 7.6 and Table 7.7.

Table 7. 6. Hierarchical model summary

<i>Model</i>	<i>Model summary</i>				<i>ANOVA</i>		<i>Change statistics</i>		
	<i>R</i>	<i>R</i> <i>square</i>	<i>Adjusted</i> <i>R</i> <i>square</i>	<i>Standard</i> <i>error</i>	<i>F</i>	<i>Sig.</i>	<i>R</i> <i>square</i> <i>change</i>	<i>F</i> <i>change</i>	<i>Sig.</i> <i>change</i>
1	0.517	0.268	0.258	0.989	28.513	0.00	0.268	28.513	0.00
2	0.536	0.287	0.273	0.979	20.808	0.00	0.019	4.222	0.04

Model 1 predictors: (constant), item length, plausible distractor

Model 2 predictors: (constant), item length, plausible distractor, text concreteness

The regression model with two item and item-text variables of item length and plausible distractors respectively was significant, $F(2, 156) = 28.513, p < .001$, and accounted for 26.8 percent of the variance in item difficulty. The adjusted R square was .258, a reduction of 1 percent if the model were generalized to the population. Adding the text variable of text concreteness into the model contributed an additional 1.9 percent of variance in item difficulty, which was statistically significant (R square change = 0.019, $F(1, 155) = 4.222, p < .05$). Therefore, model 2 with three predictor variables of item length, plausible distractors, and text concreteness was taken as the final model that best represented the prediction of item difficulty in the tests. The adjusted R square of the three predictor variables model was 0.273, indicating 1.4 percent decrease in the predictive power of the model when generalized to the population.

Table 7. 7. Parameter estimates of the hierarchical regression model

<i>Predictors</i>	<i>Unstandardized</i>		<i>Standardized</i>			<i>Collinearity</i>			
	<i>coefficients</i>		<i>coefficients</i>			<i>Correlations</i>		<i>statistics</i>	
	<i>B</i>	<i>SE</i>	<i>Beta</i>	<i>t</i>	<i>Sig.</i>	<i>Zero-order</i>	<i>Partial</i>	<i>Tolerance</i>	<i>VIF</i>
Constant	3.19	1.47		2.17	0.03				
Plausible distract.	-1.07	0.17	-0.43	-6.28	0.00	-0.45	-0.45	0.99	1.01
Item length	0.01	0.00	0.24	3.49	0.00	0.29	0.27	0.99	1.01
Text concrete.	-0.01	0.00	-0.14	-2.06	0.04	-0.18	-0.16	0.99	1.01

Plausible distractor had the most powerful standardized regression weight ($\beta = -0.43, p < .001$), followed by item length ($\beta = 0.24, p < .001$), and text concreteness ($\beta = -0.14, p < .05$), all statistically significant. Plausible distractors and text concreteness had negative β values, suggesting that one standard deviation increase in the values of these variables was conducive to an additional 0.43 and 0.14 standard deviation decrease in the values of item difficulty respectively. As item length increased by one standard deviation, item difficulty increased by an additional .24 standard deviations. Therefore, more difficult items were associated with fewer plausible distractors, less concrete texts, and longer item length. The regression equation for the final model can be rewritten as follow,

*Item difficulty = 3.19 + (-1.07*plausible distractor) + (0.01*item length) + (-0.01*text concreteness)*

7.4. Discussion

Apart from the examination of the cognitive processes while test-takers engage in the test and of the pattern of test scores that represent the underlying test structure, the meaning of the test construct also benefits from an informed understanding of the task features that determine the difficulty of the test items. This chapter contributes to the test validation project in this regard. The linguistic and discourse characteristics of the texts, items, and item-text were quantified using the automatic textual analysis tool of Coh-metrix. In a subsequent stage, the extent to which these features contribute to item difficulty was examined in correlation and regression analyses.

The correlation analysis indicated that four variables, including one text variable of text concreteness, two item variables of item length and item lexical overlap, and one item-text variable of plausible distractors had a statistically significant correlation with the test item difficulty. The identification of these salient variables and the direction of their relationship with item difficulty conformed to theoretical expectations. That is items with more plausible distractors were less difficult; items based on less concrete texts were more difficult; items with more content words in both the item stem and response options tended to be more difficult; and items with larger lexical overlap between the correct option and the distractors appeared to be more difficult. The statistically significant correlations of both text and item-text features of text concreteness and plausible distractors with item difficulty implied that test-takers engaged in processing the texts to a certain degree in answering the test questions. Surprisingly however, except for the afore-mentioned text features, none of the other text variables that were found to be significant correlates of item difficulty in previous studies were associated with item difficulty in the present study. These variables include measures of text length, syntactic complexity, cohesion, and text readability. While correlation results suggested the direction and magnitude of the relationship among the text/item variables and item difficulty, a question remains as to whether item difficulty was more a function of the item and item-text variables as opposed to the text variable of text concreteness. The latter is of more importance to reading comprehension assessments since it is the readers' understanding of the reading texts and the difficulty thereof that is assessed rather than their difficulty in comprehension due to item and item-text complexity. This was addressed in the regression analyses.

Results of the standard regression analysis suggested that except for lexical overlap, all other predictors contributed significantly to item difficulty. The strongest predictor of item difficulty was plausible distractor, followed by item length and text concreteness. Lexical overlap, though correlated significantly with item difficulty, played a minimal role in predicting item difficulty in the regression model. In the subsequent hierarchical regression analysis, the item and item-text variables of plausible distractor and item length accounted for a substantial amount of variance in item difficulty above and beyond that accounted for by text concreteness when the latter was entered in the model alone. This finding offered empirical evidence against the claim that test-takers' understanding of the reading texts was determined primarily by variables related to the text rather than those related to the test methods (items) alone. In other words, the study results introduced potential construct-irrelevant variance factors since test item difficulty was more subject to the item and item-text variables than to the reading text variables, the latter of which should be the focus of assessment though. Several useful themes emerged from the results of this phase.

Although 20 variables were originally proposed as potential predictors of item difficulty in the present study, only four of them were found to have significant correlation with item difficulty and three contributed significantly to item difficulty. The majority of proposed text features including text length, syntactic complexity, cohesion measures, and text readability seemed to be irrelevant to item difficulty of the tests. One potential explanation ensued from Rupp et al. (2006) findings on the reading behaviours of students responding to reading comprehension multiple choice items. They reported that students had a tendency to segment texts into chunks associated with a particular question and processed those chunks at microstructure rather than macrostructure levels. That the linguistic features proposed in this study were analysed at the text level might have obscured the relevant features that would have otherwise become more pronounced if considered at the individual item level. Another explanation could be that the texts included in this study might not have varied enough in terms of these textual features to enable the detection of their relationship with item difficulty. This could be attributed to the small number of reading texts and the statistical methods employed in this study that might not be sensitive enough to discern the pattern of the relationship among the text variables and item difficulty. Indeed, the use of multiple regression analysis might not have been ideal in addressing the nested structure of the item and text variables. Unfortunately, the limited number of reading texts that the researcher was able to access rendered the sample size insufficient for more robust statistical analysis procedures, such as multi-level modelling.

Plausible distractor which is an item-text variable was found to be the strongest determinant of item difficulty followed by the item variable of item length, while text concreteness, a text variable, played a less important role. The identification of these salient predictors of item difficulty lent further empirical support to Embretson and Wetzel (1987) model of cognitive processing difficulty in reading comprehension items. According to the model, the difficulty of processing reading comprehension items is governed by two components, a text representation component and a response decision component. The former refers to the encoding and retrieval of the text messages. The difficulty of this process is governed by the linguistic and discourse complexity of the texts. The latter is composed of three processes, the encoding and coherence process similar to the first component but at the item level, the text mapping process where the propositions in the response options are mapped against the information contained in the text, and the decision making process where falsification and confirmation of the response options take place. Although only three variables were found to be significant predictors of item difficulty in the current study, they represented all the components in the model. Findings implied that the test item difficulty in the current study was primarily predicted by how well students could map the propositions in the question and response options onto the information contained in the text to inform the decision-making process, and by the length of the questions and responses options. On the other hand, the text variable of text concreteness which is relevant to the text representation component only contributed minimally to item difficulty. Similar findings were also reported by Embretson and Wetzel (1987) in their validation of the model, which highlighted the significance of the response decision component in predicting item difficulty.

The finding of plausible distractors as the most powerful but negative predictor of item difficulty seemed to substantiate the argument made by Rupp et al. (2006) that students approached multiple choice questions as a problem-solving task rather than a reading comprehension task. According to them, when students approached a difficult item, they tended to shift back and forth between the response options and the relevant text portions to falsify the implausible answers until none or few of them remained. Therefore, if an item has more distractors that can be directly falsified by the information in the text, the item becomes less difficult. Another factor that might contribute to the significance of this item-text variable was the nature of the multiple-choice question per se. The type of questions that asked students to select the option that was “not mentioned in the text” or “not true according to the text” might have automatically predisposed students to resort to the strategy of eliminating implausible answers as the only pathway to find the answer. Accordingly, the number of distractors that

could be directly confirmed in the text largely determined the probability of successfully answering the item. From a practical viewpoint, the dominant effect of plausible distractor variable on item difficulty could also be explained by the “teaching to the test” instructional approach currently prevalent in Vietnam wherein eliminating implausible distractors is a frequently recommended technique to handle multiple choice test items (see Chapter IX). This was further supported by the stimulated verbal recall data presented in chapter V where the majority of interviewed students reported using this technique to answer many of the test questions. It should be noted, however, that eliminating implausible answers is associated more with general test-taking strategies than with reading comprehension skills (Ozuru et al., 2008) and therefore should be a peripheral strategy in the test-taking process to maintain the meaningfulness of the test constructs.

The significant predictor of item length introduced a construct irrelevance factor. The probability of students’ success on an item depended more on the length of the question stem and the alternatives than on the reading text per se. Items with more words in both the stem and options were more difficult for the students. This finding was consistent with what (Rupp et al., 2001) reported in their study. Potentially, longer items imposed more cognitive processing demand on students’ decoding and retrieving of item messages, thereby leaving less cognitive resources for the processing of the text and challenging the mapping of propositions in the items with the information in the text. Similarly, longer items might have also overloaded readers’ short-term memory where they temporarily stored information from the items while searching for the relevant text sections that contained the required information to answer the questions.

Finally, that the predictor variables in this study only accounted for less than a third of the variance in the item difficulty indicated that there were other variables which might manifest themselves to be more significant predictors of item difficulty but were not included in the current study. Given that the difficulty of reading comprehension items may be governed by a multitude of linguistic and discourse features at both item and text levels, there needs to be continuous exploration of these features in subsequent replication studies to shed more light on the meaningfulness of the test constructs.

In summary, this chapter presents results of the linguistic and discorsal features of the texts, items, and item-text as predictors of item difficulty in the test. Implausible answer was found to be the most robust predictor of item difficulty, followed by item length and text concreteness. These findings provide supplementary evidence for the findings in Chapter V and Chapter VI, thereby enriching the evidence for the explanation inference for validity of the

interpretation and use of the test scores. Chapter X presents a detailed discussion of these supplementary findings.

CHAPTER VIII: STUDENTS' TEST SCORES AND THEIR ENGLISH READING PROFICIENCY IN THE TARGET LANGUAGE USE DOMAIN

8.1. Introduction

This chapter reports on findings related to the extrapolation inference in the argument-based framework which makes assumptions about the relationship between students' performance on the L-VSTEP reading test and their actual performance in the target language use domain. More specifically, the research reported in this chapter aims to examine if students' scores on the L-VSTEP reading test could be used to predict their English reading performance in the academic programs that they were pursuing, as measured by their self-reported English reading proficiency via a self-assessment questionnaire. As articulated in the interpretive argument, the interpretation and use of students' scores on the L-VSTEP reading test would be meaningful to the extent that their predictive relationship with a measure of students' English reading proficiency in the target language use domain could be established. Failure to detect this relationship undermines the interpretation and use of the test scores. Out of the various potential methods that could be used to serve as the criterion measures of the L-VSTEP reading test, the English reading proficiency self-assessment questionnaire was chosen given empirical evidence in support of its use in the literature (Brantmeier & Vanderplank, 2008; Ross, 1998). The following sections, therefore, present results of the development and validation of the self-assessment questionnaire, and of the structural equation model that tests the hypothesized relationship between students' test scores and their self-reported English reading proficiency.

8.2. Validation of the self-assessment instrument

The first phase of this study sought to develop and validate a reading proficiency self-assessment scale. The scale items were developed based on a) a thorough review of the extant literature on L2 reading comprehension which was discussed in chapter II, b) the CEFR-VN, c) the guidelines for the development of the L-VSTEP reading test, d) the L2 reading comprehension curriculum adopted at the institution where the data collection took place, and e) DeVellis (2016) suggestions for best practices in scale construction and validation. Detailed methodology for instrument development has already been reported in chapter IV.

8.2.1. Exploratory factor analysis

Descriptive statistics and univariate normality check

The data set consisting of 344 participating students was initially subjected to Exploratory Factor Analysis (EFA), using IBM SPSS 22 software, to identify the underlying factor structure of the self-assessment scale (Field, 2009; Hair et al., 2014; Loewen & Gonulal, 2015). Table 8.1 presents descriptive statistics pertaining to students' responses to the self-assessment questionnaire.

Table 8. 1. Descriptive statistics of the self-assessment questionnaire (N = 344)

Items	Mean	SD	Skewness	Kurtosis
1	4.72	.96	-.91	1.42
2	4.40	.98	-.41	.17
3	4.70	.82	-.45	.25
4	4.43	.84	-.52	.98
5	4.78	.96	-1.02	1.61
6	4.52	.92	-.51	.65
7	3.52	.97	-.27	-.23
8	3.99	.88	-.37	.54
9	3.81	.95	-.38	.28
10	4.12	.97	-.40	-.09
11	4.16	.91	-.47	-.04
12	4.26	.87	-.59	.41
13	3.97	.96	-.07	-.29
14	4.10	.88	-.18	.49
15	4.14	.94	-.52	.23
16	3.57	.91	-.29	.11
17	3.70	.96	-.07	-.17
18	3.27	1.11	-.002	-.41
19	3.81	.91	-.21	-.25
20	3.99	.88	-.34	.20
21	3.80	1.07	-.43	-.27
22	2.93	1.04	.09	-.502
23	2.91	1.09	.22	-.074
24	2.86	1.12	.29	-.172
25	3.42	1.08	-.23	-.324
26	2.71	1.06	.16	-.438

27	3.25	1.17	-.09	-.568
----	------	------	------	-------

The mean scores of the 27 questionnaire items ranged from 2.71 to 4.78 with standard deviations in the range of 0.87 to 1.17. The largest absolute values of skewness and kurtosis were |1.02| and |1.61| respectively, both lower than the cut-off values of |3| and |8| as signals of departure from univariate normality. To determine if the data were suitable for EFA, the assumptions of sampling adequacy and variable correlation were checked, using the Kaiser – Meyer – Olkin measure of sampling adequacy (KMO) (Kaiser, 1970) and the Barlett’s Test of Sphericity respectively. The KMO value was .896 indicating that the pattern of correlation among items was compact, and therefore distinct and reliable factors could be extracted (Field, 2009). The Barlett’s test of sphericity was significant, $\chi^2(351) = 3625.493$, $p < .001$, suggesting that the correlations among items were sufficiently large for EFA.

Since theoretical premises and empirical research were indicative of the relatedness of L2 reading subskills, principal axis factor analysis with promax rotation and Kaiser normalization was conducted on the self-assessment data. Six components with eigenvalues above 1 were initially yielded and cumulatively accounted for 57.8 percent of the variances. The scree plot as exhibited in Figure 8.1 was relatively hard to interpret. The curve tailed off after the third factor and then slightly dropped after the sixth factor. Given the relatively large sample size in the study, it was safe to consider the Kaiser’s eigenvalues as a reliable criterion in determining the number of factors.

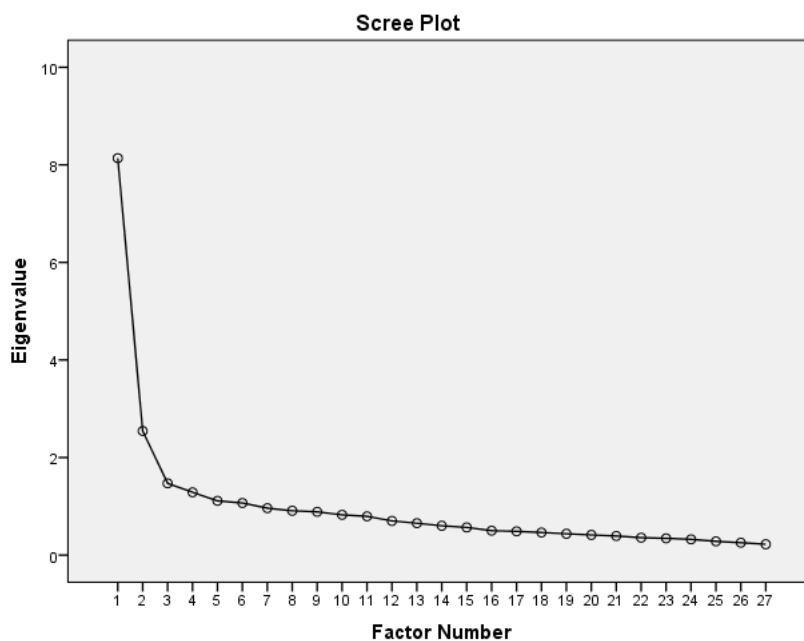


Figure 8. 1. The scree plot

Factor rotation maximized the loadings of items on one factor while minimizing their loadings on the remaining factors (Loewen & Gonulal, 2015), thereby enhancing the interpretability of the extracted factors. This process yielded six factors as in the unrotated solution stage, all having their corresponding items with a factor loading above .40. As suggested by Field (2009), factor loadings below .40 which might undermine the substantive importance of a variable to a factor were suppressed while variables that had high cross-loadings on more than one factor were excluded from the final factor solution. As a result, six items were removed. Table 8.2 presents the factors (reading subskills) that were retained, their associated items, and relevant factor loadings.

Table 8. 2. Factor solution and factor loadings

Factors	Items	Item content	Factor loadings
Understanding Explicit Information	UEI1	I can understand factual information in a text, using grammatical structures and vocabulary familiar to me.	.683
	UEI2	I can understand an explicit detail in the text, but rewritten using different words.	.651
	UEI3	I can understand concepts/ideas in sentences that use familiar grammatical structures and vocabulary, have familiar topics and clear organization.	.534
Understanding Pragmatic Meaning	UPM1	I can identify the perspectives and stances of the text's author.	.553
	UPM2	I can understand the tone of the author in a text.	.493
	UPM3	I can identify the message that the author wants to convey via the text.	.758
	UPM4	I can understand the purpose of the author given a detail in the text.	.501
	UPM5	I can understand the logical inferences/arguments of the author.	.546
	UCD1	I can identify the antecedent of a pronoun.	.623

Understanding Cohesive Devices	UCD2	I can understand the logical ideas among sentences in a text, based on reference words, linking words, connectives, or repeated words.	.551
Summarizing	STI1	I can understand the main idea of a whole text.	.449
Textual Information	STI2	I can understand the main idea of a paragraph.	.763
	STI3	I can draw the conclusions from a passage.	.682
Integrating	ITI1	I can integrate information across the text to establish the logical connection among ideas.	.667
Textual Information	ITI2	I can integrate information in the text with my prior knowledge to understand an argument/detail.	.647
Inferring	ISM1	I can infer the meaning of colloquial or idiomatic expressions from the context.	.739
Situational Meaning	ISM2	I can infer the implied meaning of a detail based on information within the text.	.706
	ISM3	I can infer the implied meaning of an argument in the text.	.777
	ISM4	I can infer the meaning of a subtle detail about an opinion/inference/attitude in a text.	.684
	ISM5	I can infer the implied meaning of a sentence/detail in a text using my prior knowledge.	.884
	ISM6	I can infer word meanings from contexts (words with different meanings)	.738

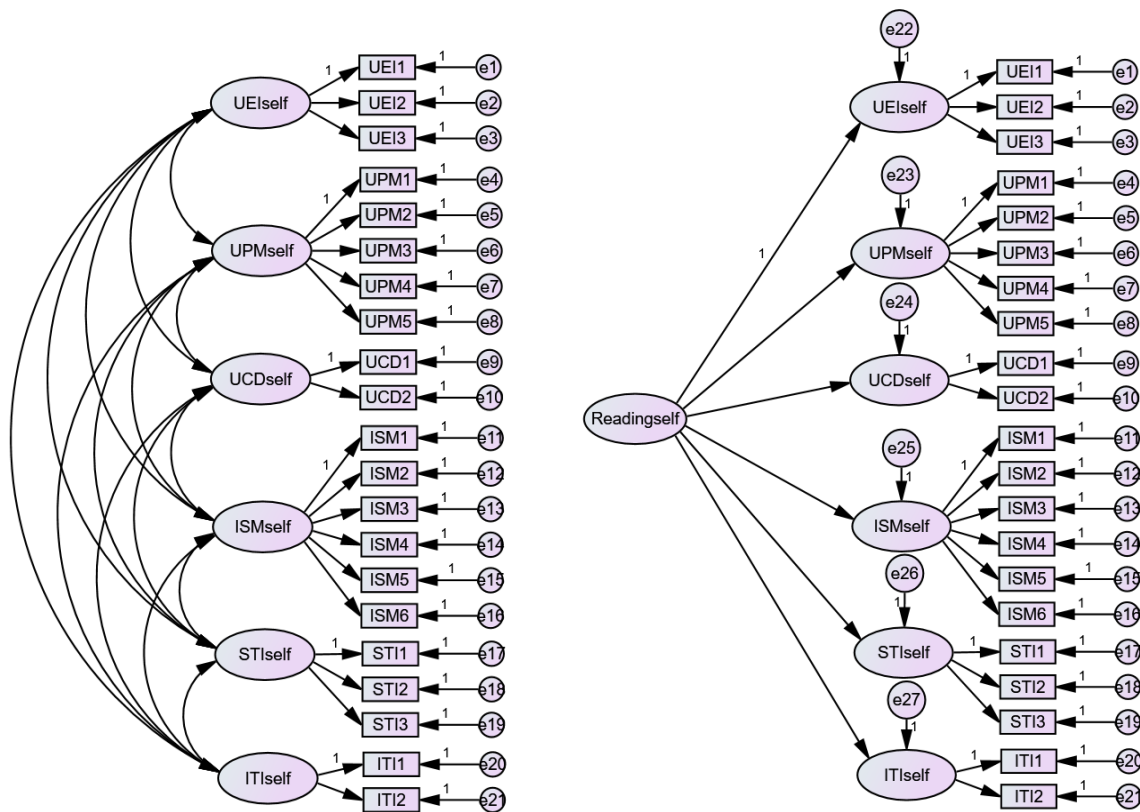
After all, six factors with 21 items were retained. These factors were labelled as understanding explicit information (UEI), understanding pragmatic meaning (UPM), understanding cohesive devices (UCD), summarizing information (STI), integrating information (ITI), and inferring information (ISM). Understanding explicit information includes three items denoting students' ability to understand factual information, details, or concepts explicitly stated in the texts. Understanding pragmatic meaning is composed of five

items which indicate students' ability to decipher the author's purposes, perspectives, stance and attitude in the text. Understanding cohesive devices contains two items indicative of students' ability to understand the logical flow of ideas in the text based on cohesive devices such as pronouns, linking words, and repeated words. Summarizing information consists of three items that demonstrate students' ability to understand the main ideas of a text either at the paragraph or inter-textual level. Integrating information includes two items pertaining to students' ability to synthesize information either across the text or with their prior knowledge to understand the logical connection among ideas in the text. Finally, inferring information comprises six items indicating students' ability to make inferences based on information within the text or their prior knowledge.

8.2.2. Confirmatory factor analysis

The six factors extracted were subsequently subjected to a confirmatory factor analysis, using the same data set of 344 students. In line with current conceptualizations of L2 reading comprehension, relevant L2 reading research as reviewed in chapter III, and the examination of the factor structure of the L-VSTEP reading test reported in chapter VI, two confirmatory factor analytic models were proposed for the self-assessment data, namely a correlated six-factor model and a second-order factor model. The one-factor model was not considered since the EFA results suggested that students' responses to the self-assessment questionnaire could be better explained by six factors of reading subskills rather than only one factor.

The correlated six-factor model (Model A in Figure 8.2) posited that students' responses to the self-assessment questionnaire could be represented by six correlated factors of reading subskills. This model is compatible with the view that L2 reading is a multidimensional construct consisting of a number of discernible but correlated subskills. The second-order factor model (Model B in Figure 8.2) is similar to the correlated six-factor model except that the correlation among the six subskills of L2 reading is now fully explained by a hypothetical second-order factor of L2 reading proficiency. This model is consistent with the view that L2 reading proficiency is a general construct with several lower-level discernible reading subskills.



Model A: Correlated six-factor model

Model B: Second-order factor model

Figure 8. 2. The hypothesized CFA models

Results

While the skewness and kurtosis values were within the acceptable range for univariate normal distribution, the assumption of multivariate normality was violated, evidenced by the Mardia’s normalized estimate of 31.31, far beyond the suggested value of 5 (Kline, 2016). Therefore, the Maximum Likelihood estimation method using Bollen-Stine bootstrapping procedure was adopted for the CFA analyses. The global model fit indices for each model are presented in Table 8.3.

Table 8. 3. Goodness-of-fit indices for the three CFA models

Goodness-of-fit indices	Correlated factor model (Model A)	Second-order factor model (Model B)
χ^2	406.690	441.008
p	.000	.000
χ^2/DF	2.351	2.423
CFI	.914	.904
TLI	.895	.890
$SRMR$.057	.065

<i>AIC</i>	522.690	539.008
<i>BIC</i>	745.447	727.200
<i>RMSEA</i>	.063	.064
<i>RMSEA CI</i>	.055 - .071	.057 - .072

*CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean-Square Residual; RMSEA = Root Mean Square Error of Approximation

*Good model fit is indicated by non-significant χ^2 , normed χ^2 (χ^2/DF) < 3, CFI and TLI > 0.90, SRMR < 0.07, SRMEA < 0.08, and narrow RMSEA confidence intervals associated with a RMSEA value below .08.

The goodness-of-fit indices for the correlated six-factor model were reasonably good ($\chi^2/DF = 2.35$, $p < .001$; SRMR = .057; CFI = .914; TLI = .895; RMSEA = .063, CI [.055; .071]). The standardized parameter estimates of the correlated six-factor model are presented in the diagram in Figure 8.3. The standardized factor loadings for each congeneric set of variables were moderate to high, ranging from .40 (UPMself \rightarrow UPM1) to .83 (ISMself \rightarrow ISM5), all statistically significant at $p < .001$ level. Each latent component of L2 reading subskills explained a relative moderate but statistically significant amount of variances in their associated observed indicators. All the latent components were significantly correlated with one another, though the magnitude of the correlations varied widely. The lowest correlation ($r = .14$) was between students' self-assessment of their ability to understand explicit meaning (UEI) and to infer situational meaning (ISM), while the largest correlation ($r = .71$) was between their self-assessment of the ability to understand cohesive devices (UCD) and to summarize textual information (STI). Note that there was an additional specification of the covariance between two error terms e14 and e16, as suggested by the modification indices in the initial run of the CFA model. This covariance was statistically significant with a medium magnitude of .32.

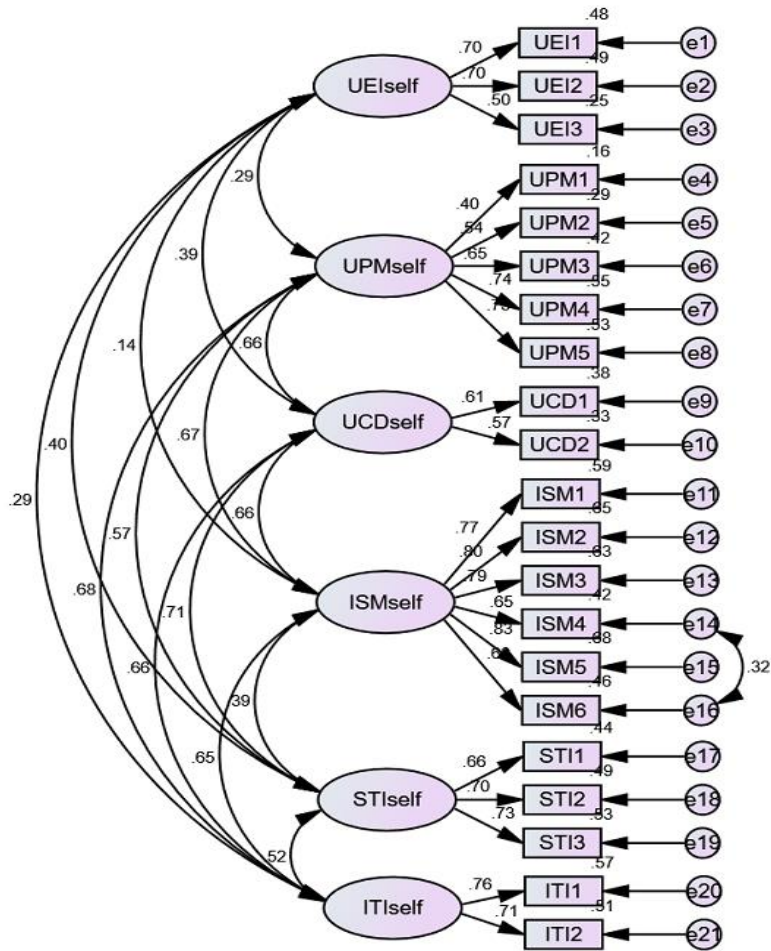


Figure 8.3. The standardized parameter estimates of the correlated factor model

The second-order factor model produced acceptable model fit, though not as good as the correlated factor model ($\chi^2/DF = 2.42, p = < .001$; SRMR = .065; CFI = .904; TLI = .890; RMSEA = .064, CI [.057; .072]). Except for the standardized factor loading of understanding explicit information (UEI) which was medium, other standardized loadings at the higher order level were large, ranging from .65 (STI) to .86 (UCD), all being statistically significant at $p < .001$ level.

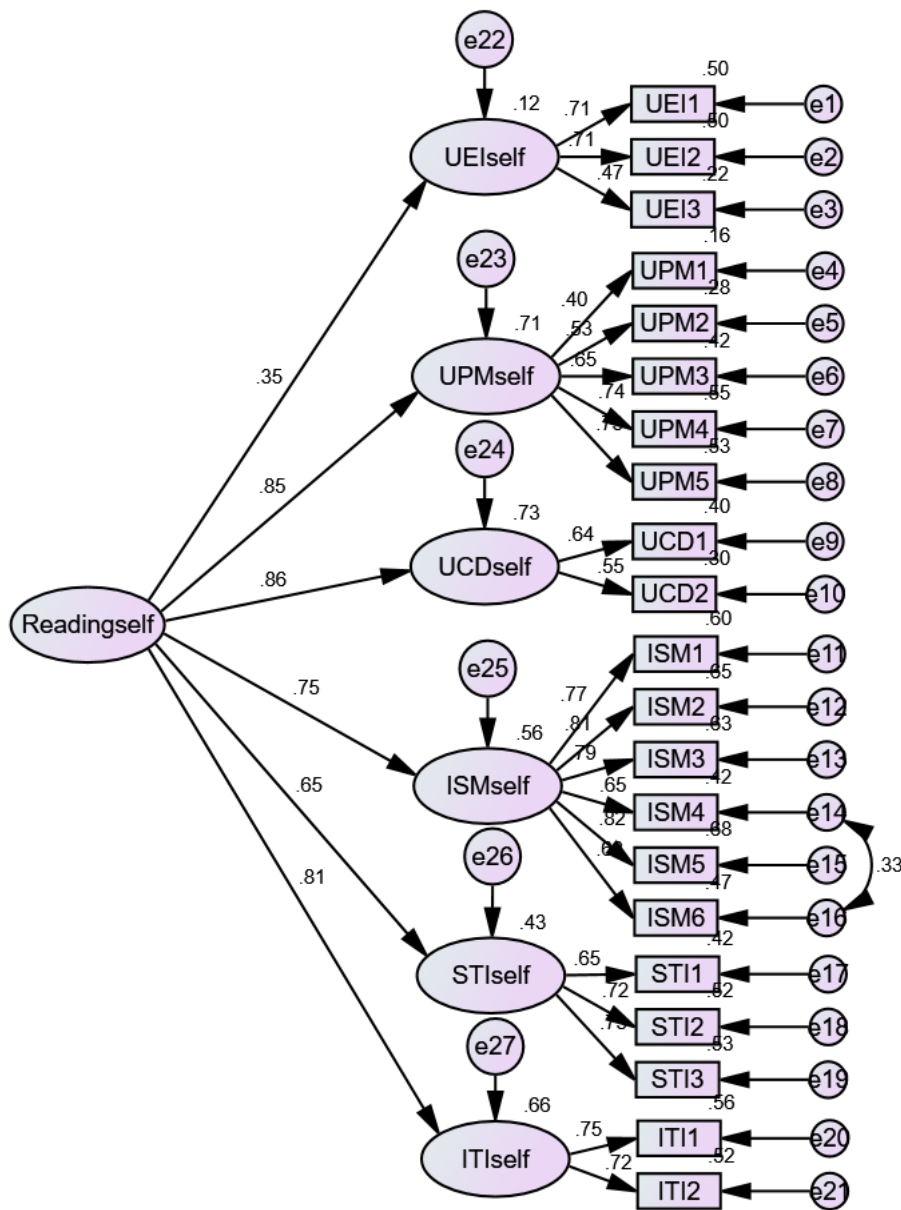


Figure 8. 4. Standardized parameters of the second order factor model

As shown in Figure 8.4, all standardized factor loadings for each congeneric set of variables at the first order level were statistically significant, with relatively comparable magnitudes to those found in the correlated factor model. Similarly, the amount of variance in each observed indicator accounted for by their corresponding latent components was all statistically significant and comparable to those reported in the correlated factor model. The additional covariance between two error terms e14 and e16 was also significant with medium magnitude of .33. Given that the second order factor model produced largely comparable model fit and parameter estimates to the correlated six factor model and that the former (df = 182) was more parsimonious than the latter (df = 173), the second order factor model was chosen as the final model representing students' patterns of responses to the self-assessment

questionnaire. Table 8.4 presents the unstandardized parameter estimates, standard errors of measurement, critical ratios and probability values of the final chosen model.

Table 8. 4. Unstandardized parameter estimates of the second-order factor model

<i>Regression weight</i>	<i>Unstandardized Estimate</i>	<i>S.E.</i>	<i>C.R.</i>	<i>p</i>	<i>Standardized estimates</i>
Reading → UEIself	1.000				.349
Reading → UPMself	1.204	.309	3.899	***	.845
Reading → UCDself	2.213	.513	4.318	***	.857
Reading → ISMself	2.546	.573	4.446	***	.749
Reading → STIself	1.760	.421	4.178	***	.655
Reading → ITIself	2.352	.533	4.415	***	.815
UEIself → UEI1	1.000				.706
UEIself → UEI2	1.027	.147	6.996	***	.709
UEIself → UEI3	.674	.104	6.458	***	.474
UPMself → UPM1	1.000				.401
UPMself → UPM2	1.403	.234	6.001	***	.533
UPMself → UPM3	1.834	.283	6.473	***	.646
UPMself → UPM4	2.111	.313	6.748	***	.741
UPMself → UPM5	2.410	.359	6.713	***	.727
UCDself → UCD1	1.000				.636
UCDself → UCD2	.850	.117	7.269	***	.552
ISMself → ISM1	1.000				.774
ISMself → ISM2	1.093	.070	15.522	***	.805
ISMself → ISM3	1.111	.073	15.246	***	.793
ISMself → ISM4	.868	.072	12.009	***	.645
ISMself → ISM5	1.093	.068	15.957	***	.825
ISMself → ISM6	.996	.078	12.825	***	.683
STI self → STI1	1.000				.651
STI self → STI2	1.031	.107	9.652	***	.719
STI self → STI3	1.001	.103	9.691	***	.726
ITIself → ITI1	1.000				.747
ITIself → ITI2	.932	.092	10.183	***	.723

Variiances

Reading	.055	.024	2.331	.020	
e22	.401	.077	5.204	***	
e23	.032	.011	2.915	.004	
e24	.098	.049	1.988	.047	
e25	.281	.043	6.468	***	
e26	.229	.044	5.170	***	
e27	.155	.040	3.858	***	
e1	.460	.070	6.592	***	
e2	.477	.073	6.514	***	
e3	.716	.063	11.440	***	
e4	.586	.047	12.571	***	
e5	.558	.046	12.002	***	
e6	.527	.048	11.099	***	
e7	.411	.043	9.663	***	
e8	.583	.059	9.946	***	
e9	.546	.039	8.449	***	
e10	.609	.040	10.396	***	
e11	.430	.039	10.912	***	
e12	.416	.040	10.386	***	
e13	.468	.044	10.612	***	
e14	.676	.056	12.000	***	
e15	.360	.036	9.962	***	
e16	.725	.062	11.773	***	
e17	.544	.054	10.147	***	
e18	.397	.045	8.734	***	
e19	.362	.042	8.583	***	
e20	.367	.047	7.766	***	
e21	.366	.043	8.448	***	
Covariance					
e14 <--> e16	.228	.045	5.096	***	.325

* Parameters are significant at .05 level; ** Parameters are significant at .01 level; ***

Parameters are significant at .001 level

8.3. Test performance and self-reported English reading proficiency

The extent to which students' performance on the test can predict their perceived performance in the target language use domains was captured by the structural equation model in which students' test data was regressed onto their self-assessment data. The former represents their immediate performance on the test while the latter represents their self-reported performance in the target language use domain. In order to testify if the factor structure of students' test data identified in chapter VI (N = 544) could be reproduced in the sample that was used in the current chapter (N=344), the established one-factor CFA model of students' test performance in chapter VI was performed again on the sample of 344 students who both took the test and responded to the self-assessment questionnaire. This model yielded exceptionally good fit ($\chi^2/DF = 1.005$, $p = .444$; SRMR = .029; CFI = 1.000; TLI = .999; RMSEA = .004, CI [.000; .052]). All parameter estimates of the test model, as presented in Table 8.5, were significantly different from zero, rendering all measurement properties of the model suitable for the structural equation model.

Table 8. 5. Parameter estimates of the test model

Weight	Estimate	S.E.	C.R.	<i>p</i>	Error	Estimate	S.E.	C.R.	<i>p</i>
UEItest	1.000				e1	.021	.002	8.967	***
LItest	.780	.139	5.604	***	e2	.038	.003	11.743	***
UCDtest	.973	.147	6.624	***	e3	.031	.003	10.452	***
ITItest	.754	.119	6.330	***	e4	.023	.002	10.970	***
ISMtest	.980	.149	6.593	***	e5	.032	.003	10.517	***
UPMtest	.463	.133	3.465	***	e6	.047	.004	12.719	***
STItest	.503	.173	2.915	***	e7	.082	.006	12.841	***
					Reading	.013	.003	4.973	***

* Parameters are significant at .05 level; ** Parameters are significant at .01 level; ***

Parameters are significant at .001 level

The structural equation model of reading test and reading self-assessment data yielded reasonably acceptable fit ($\chi^2/DF = 1.823$, $p < .001$; SRMR = .059; CFI = .905; TLI = .895; RMSEA = .049, CI [.043; .055]). All parameters, including factor loadings and variances, in the model were statistically significant. As shown in Figure 8.6, except for the two variables of UPMtest and STItest, the standardized factor loadings of all other observed variables onto their corresponding latent constructs ranged from medium (.36) to strong (.85). The standardized factor loadings of UPMTest (understanding pragmatic meaning) and STItest

(summarizing textual information) onto the reading test construct were rather modest, only .25 and .21 respectively. The standardized regression coefficient from the reading test construct to the reading self-assessment construct was statistically significant with a magnitude of .35, meaning that as students' test scores increase or decrease by one standard deviation, their self-assessment scores change by 0.35 standard deviations accordingly. The amount of variance in the reading self-assessment measure accounted for by the reading test measure was 12 percent, suggesting that a large amount of variance in students' self-assessment (88 percent) was unexplained. The unstandardized factor loadings, standard errors of measurement, critical ratio values and probability values of the model are presented in Table 8.5.

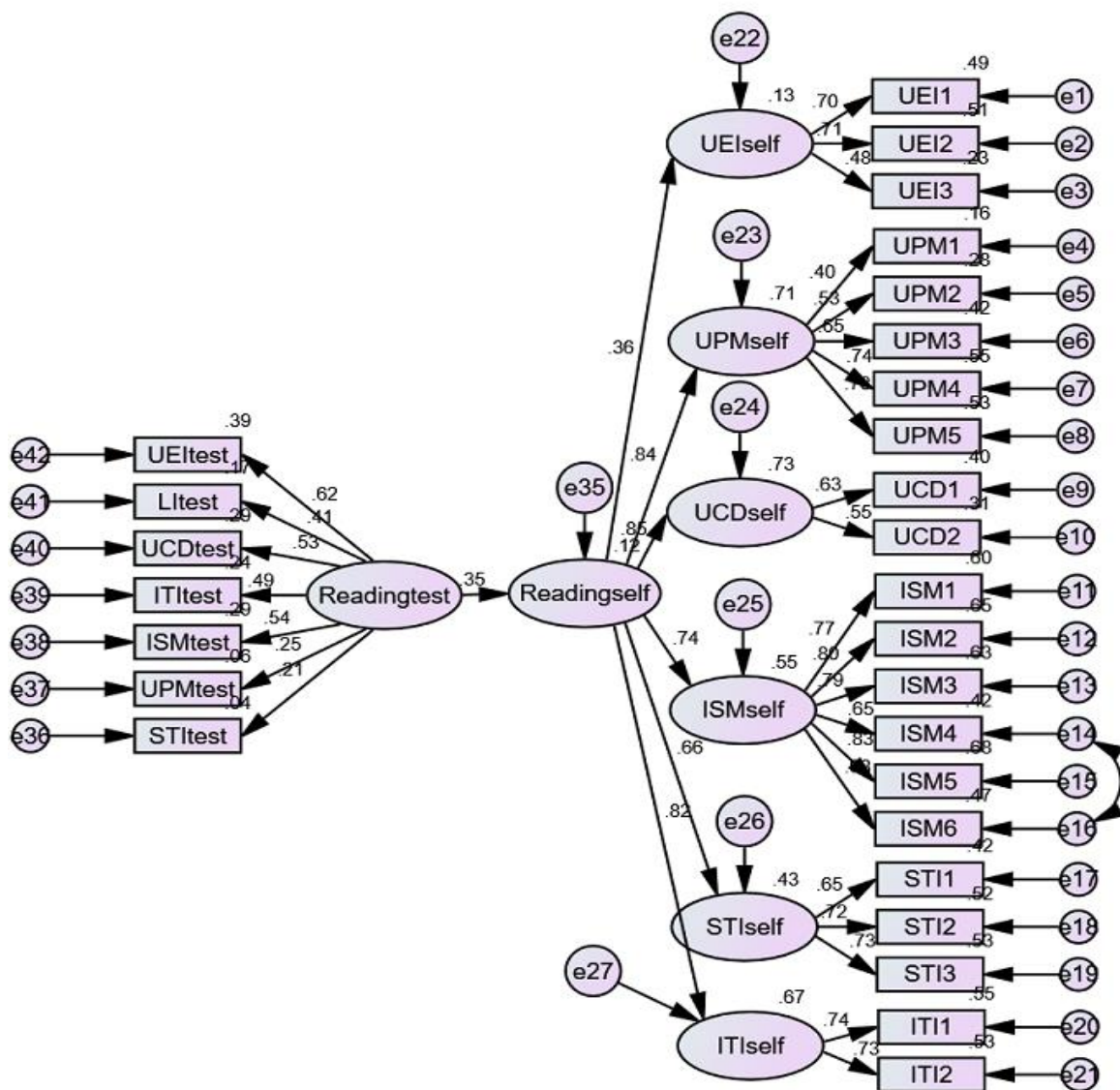


Figure 8. 5. The structural equation model

Table 8. 6. The unstandardized parameter estimates of the SEM model

Weights	Estimates	S.E.	C.R.	<i>p</i>	Errors	Estimates	S.E.	C.R.	<i>p</i>
Path	1.404	.590	2.380	.017	e1	.465	.069	6.768	***
UEIself	1.000				e2	.473	.073	6.499	***
UPMself	1.166	.290	4.025	***	e3	.715	.063	11.432	***
UCDself	2.114	.473	4.469	***	e4	.584	.046	12.562	***
ISMself	2.426	.526	4.612	***	e5	.557	.046	12.001	***
STIself	1.701	.393	4.325	***	e6	.527	.047	11.101	***
ITIself	2.264	.494	4.585	***	e7	.413	.043	9.700	***
UEI1	1.000				e8	.583	.059	9.947	***
UEI2	1.037	.147	7.070	***	e9	.549	.064	8.509	***
UEI3	.681	.105	6.500	***	e10	.607	.059	10.350	***
UPM1	1.000				e11	.431	.039	10.919	***
UPM2	1.394	.231	6.040	***	e12	.418	.040	10.404	***
UPM3	1.821	.279	6.520	***	e13	.468	.044	10.603	***
UPM4	2.090	.308	6.796	***	e14	.676	.056	11.999	***
UPM5	2.392	.354	6.765	***	e15	.358	.036	9.932	***
UCD1	1.000				e16	.724	.062	11.769	***
UCD2	.856	.118	7.271	***	e17	.545	.054	10.162	***
ISM1	1.000				e18	.397	.045	8.752	***
ISM2	1.093	.071	15.479	***	e19	.361	.042	8.594	***

ISM3	1.113	.073	15.236	***	e20	.369	.047	7.856	***
ISM4	.868	.072	12.001	***	e21	.364	.043	8.436	***
ISM5	1.095	.069	15.961	***	e22	.391	.075	5.205	***
ISM6	.997	.078	12.822	***	e23	.033	.011	2.945	**
STI1	1.000				e24	.099	.049	2.019	*
STI2	1.031	.107	9.663	***	e25	.287	.044	6.552	***
STI3	1.002	.103	9.693	***	e26	.227	.044	5.163	***
ITI1	1.000				e27	.153	.040	3.842	***
ITI2	.936	.091	10.237	***	e35	.053	.022	2.420	*
STItest	1.000				e36	.082	.006	12.827	***
UPMtest	.929	.371	2.501	*	e37	.047	.004	12.688	***
ISMtest	1.885	.620	3.042	**	e38	.031	.003	10.528	***
ITItest	1.398	.466	2.999	**	e39	.023	.002	11.181	***
UCDtest	1.829	.602	3.037	**	e40	.031	.003	10.635	***
LItest	1.450	.499	2.907	**	e41	.038	.003	11.860	***
UEItest	1.885	.611	3.083	**	e42	.021	.002	9.219	***

* Parameters are significant at .05 level; ** Parameters are significant at .01 level; *** Parameters are significant at .001 level

In a nutshell, the fairly good model fit and appropriate parameter estimates of the structural equation model suggested a predictive relationship between students' scores on the test and their self-reported performance in the target language use domains as captured by the self-assessment questionnaire.

8.4. Discussion

As articulated in the interpretive argument, the explanation inference addresses the link between the L-VSTEP reading test's expected scores and the theoretical construct of L2 reading proficiency that underlies the development of the test. The extrapolation inference, by extension, offers a bridge that links the L2 reading proficiency construct to the target scores which are claims about the quality of students' performance in the target language use domains. A difficulty with which relevant backings could be sought to warrant these claims is the lack of a viable criterion measure that can serve as a representation of students' performance in the target language use domains. Several measures have been proposed and testified in previous studies, among which the use of a self-assessment questionnaire to record students' self-perception about their current language abilities received both theoretical and empirical justification (Brantmeier & Vanderplank, 2008; Fan & Yan, 2017; Li, 2015b; Ross, 1998). This chapter, therefore, started with the development and validation of a self-assessment questionnaire to capture students' L2 reading proficiency in the target language use domains, drawing on the CEFR-VN, the guidelines for the L-VSTEP reading test item writing, and the curriculum for English learning at the institution where the study was conducted. The self-assessment data were then employed in a structural equation model that depicts the hypothesized relationship between students' test scores and their self-assessment responses.

Results of the EFA extracted six factors of L2 reading subskills as measured via students' responses to the questionnaire. These include students' ability to understand information explicitly stated in the text, to understand authors' purposes, positions, and attitude, to understand logical connection among ideas based on connective devices, to infer information from the text, to summarize information in the text, and to integrate information across the text. These subskill components were compatible with those found via students' test scores as reported in chapter V except that the ability to guess the meaning of an unfamiliar word was now subsumed in the ability to infer information from the text. This is conceivable given that the ability to infer the meaning of an unfamiliar word can be likened to the ability to make inferences based on information in the text (i.e. the surrounding context of the word) and on students' prior knowledge (vocabulary knowledge), but at the lexical level.

Two confirmatory factor analytic models that capture the underlying relationship among the components derived from EFA were tested via confirmatory factor analysis. Results supported a second-order factor model wherein students' responses to the questionnaire could be explained by a general L2 reading comprehension component which was divisible into discernible sub-components. This finding corresponds with the view that L2 reading is a multidimensional construct which informs the development of L2 reading instructional approaches and assessment practices across different contexts (Alderson, 2000; Song, 2008). The finding of a meaningful underlying structure of the self-assessment data lent further empirical support to previous studies that found self-assessment as a viable learner-directed criterion measure of their language proficiency in the target language use domains (Fan, 2016; Li, 2015b). What made this study stand out from previous research was the focus of the self-assessment on a macro-skill of language ability rather than on language ability as an overarching construct. The development and scrutiny of questionnaire items were rather challenging due to the level of detail required to craft each item and the need to maintain the consistency of the item content as informed by the Framework, the test specifications, and the curriculum. The challenges were also evident in the piloting phase wherein students had difficulty understanding the subtle differences among items that were designed to assess the same subskills, which necessitated several rounds of item wording and revision. The study findings, therefore, suggested that to the extent that the items were carefully crafted based on both theoretical and empirical justification, a self-assessment questionnaire could have the potential to capture students' self-perceived macro-language skill proficiency as reliably as the self-assessment instrument for language ability as a whole.

Except for the understanding explicit information factor, the other five factors of self-assessed English reading proficiency loaded highly on the general construct of L2 reading, with magnitudes ranging from .65 to .86. The factor loading of the understanding explicit information factor was relatively modest (.36) and only 12 percent of its variance was accounted for by the higher-order construct of L2 reading. This could be because the number of items designed to target this construct were not sufficiently representative to elicit meaningful and reliable responses from the students. Adding more items targeting this factor might increase the explanatory variance of the construct in future studies. Another possible explanation was that students might not have accurately judged their reading ability at this particular subskill level. As argued by Powers and Powers (2015), one of the drawbacks of self-assessment was that it had a propensity to make students succumb to the temptation to overestimate their skills, thereby presenting themselves in "socially desirable ways" (p.157).

Since participants in this study were students in their final year at university who were assumed to have completed all language skill modules and hence should have achieved at least some basic skills in reading, the inclusion of items to assess this low-level reading subskill in the questionnaire might have prompted them to supply unduly biased ratings.

The structural equation model was proposed and tested to yield more insights into the relationship between students' scores on the L-VSTEP test and their responses to the self-assessment questionnaire items. The generally acceptable model fit and statistically significant parameter estimates were indicative of the appropriateness of the proposed model to capture the relationship between students' scores on the test and their self-assessment responses. As students' test scores changed by one standard deviation, their respective self-assessment scores changed by additional .35 standard deviations, a relatively moderate regression coefficient. This regression weight was somewhat less optimum than that reported in Fan and Yan (2017) study (.52) which was conducted at the language ability level rather than at the macro-skill level of reading. Yet, given the exploratory nature of this study, the statistically significant and moderate regression weight appeared to be promising and suggested that the link between students' target scores on the L-VSTEP reading test and their self-reported performance in the target language use domain could be established with relative confidence. This, again, augmented the earlier argument that the self-assessment questionnaire could be used as a potential criterion measure in predictive validity studies of language tests. However, similar to the study reported by Fan and Yan (2017), only a modest amount of variance (12 percent) in the self-assessment factor was accounted for by the test factor, leaving a significant amount of variance (88 percent) unexplained. This could be attributed to the various individual characteristics and experiential factors that might affect the self-assessment practice as suggested in previous studies (Alderson, 2005; AlFallay, 2004; Liu & Brantmeier, 2019; Suzuki, 2015).

Suzuki (2015), for example, found that the amount of exposure to reading materials and the length of residence in the naturalistic acquisition context affected the accuracy of students' self-assessment of their reading ability. A similar phenomenon might have been observed in the current study as students were recruited from different academic disciplines who were likely to be exposed to different reading materials with different text types, topics, and genres, taught by different teachers, and followed different learning curricula.

Similarly, Brantmeier and Vanderplank (2008) reported that the level of accuracy of students' self-assessment of their reading ability varied with respect to the different task types, such as written recall, multiple choice items, and sentence completion included in the reading

tests. That the students in the current study might have encountered a variety of different reading task types during their learning process while all items in the L-VSTEP reading test had a multiple choice response format might offer some plausible explanations for the vast amount of unexplained variance in students' self-assessment of their reading ability.

Another potential confounding variable that might affect the accuracy of self-assessment is the levels of training and practice required of students to be able to confidently and accurately judge their own ability (Dolosic, Brantmeier, Strube, & Hogrebe, 2016; Goto Butler & Lee, 2010). Due to the cross-sectional and large-scale nature of the current study, extensive training programs to develop students' self-assessment skills prior to data collection was, therefore, not conducted, which might raise some concerns about the extent to which students accurately self-assessed their skills. Even though a link between students' test scores and their self-assessed reading ability could be reliably established via the statistical modelling in this study, the accuracy of self-assessment could have been leveraged had a more extensive and thorough training program been undertaken. Likewise, future studies that set the accuracy of self-assessment as the central focus should take the multitude of learner, environment, and task variables discussed above so as to facilitate and uphold the accuracy of students' self-assessment of their ability.

Although the measurement properties of both the reading test and reading self-assessment CFA models were adequate for structural equation modelling, the factor loadings of the understanding pragmatic meaning subskill (UPMtest) and summarizing textual information subskill (ISMtest) onto the latent construct of reading test were rather low, with magnitudes of .25 and .21 respectively. Recall that these two factor loadings were found in chapter 6 to be non-invariant across students with different academic disciplines. This might also contribute to the large residuals in students' self-assessment yielded as a result of the structural equation modelling approach. This finding underscored the importance of testifying the measurement invariance properties of both the tests and the criterion measure of self-assessment before further analysis could be conducted, particularly in studies where participants had diverse demographic and academic backgrounds. It follows, therefore, as an implication of this study that if self-assessment instruments are to be used as low-stakes, temporary measures of students' reading ability for formative purposes, the development of self-assessment instruments should be tailored to students with different demographic and academic backgrounds, to different task types and individual characteristics.

In sum, results of this chapter generally support the extrapolation inference which make claims about the relationship between students' performance on the test and their perceived

performance in the target language use domain as captured by the self-assessment questionnaire. However, the large standardized residuals left after the structural equation modelling raised some concerns about the extent to which students' self-assessed reading ability could be accurately predicted by their scores on the test, which in turns presented rebuttals against the interpretative argument as laid out in chapter four. Further discussion of these validity issues will be provided in the evaluation of the validity argument in chapter X.

CHAPTER IX: ALIGNMENT OF THE L-VSTEP READING TEST TO THE TARGET LANGUAGE USE DOMAINS

9.1. Introduction

This chapter reports on findings related to the comparability between the reading test tasks and skills included in the L-VSTEP reading test and those commonly encountered in the target language use domains which are the academic language learning environments of the students. These findings contribute to a more informed understanding of the extent to which students' scores on the test could be generalized beyond the test to the target language use situations where students' actual L2 reading occurs. As laid out in the interpretive argument in chapter IV, the extrapolation inference makes a claim about the relationship between students' performance on the test and their performance in the target language use situation. This claim can be premised on several warrants, one of which is the degree to which reading tasks and skills required in the target language use situation, herein the academic language learning environment of the students, are represented in the test. Unless there is a close alignment between the reading tasks and skills presented in the L-VSTEP reading test and those that are considered important and commonly employed or taught in the relevant academic settings, the extrapolation of students' test scores beyond the test per se and into the real-world domains will be compromised. Therefore, this alignment or lack thereof determines the types of evidence adduced to support or rebut the claim made about students' performance on the test. To that end, semi-structured interviews were conducted with both lecturers who are teaching English and graduates who have just completed their Bachelor programs and the L-VSTEP test at the institution where the study was based. The interview data and the emerging findings revolve around the alignment between the reading tasks and skills assessed in the test and those required in the real-world academic programs at the same institution, thereby offering insights into the relevant stakeholders' perceptions about the alignment between the test and the academic settings. Results of this phase of the project are presented in the following sections.

9.2. Findings

The presentation of findings in this chapter revolves around four key themes that emerged from the interview data, including the perceived importance of English reading skills in the academic programs, the amount and types of reading required in the academic programs, the reading tasks and skills required in the academic programs, and the perceived comparability

of reading tasks and skills required in the academic programs and those found in the L-VSTEP reading test.

9.2.1. The perceived importance of English reading in the academic programs

A general theme that emerged from both the lecturers and students' accounts was the importance of English reading ability in both the academic programs and the future working environment, the latter of which, though not directly related to the main aspect under investigation, was repeatedly mentioned by the participants. Of note was the perceived importance of English reading ability to English major students either in the English teaching program or English for Translation and Interpretation program. For English major students, reading is considered an essential skill, the lack of which may seriously affect their performance in the academic programs, as explained by Lecturer B.

“The ability to read in English is particularly crucial because all the prescribed materials in their programs are in English.... And their homework as well, if they cannot read, they cannot do anything. These are basic needs of English reading. But when they are in second, third, and final year, the importance of reading increases because they need to read longer, more complicated materials in other subjects such as English literature and culture, teaching methodology, so on.”

As revealed in the quote above, the importance of English reading is not only to the reading module which is exclusively focused on the training of reading skills, but also in other modules wherein all the prescribed materials are delivered in English. The importance of English reading increases as students advance through their degrees where the majority of other modules expose them to extended English reading materials. In addition, the specialized English programs which students are following also predispose them to perform certain learning tasks in which the ability to read English materials plays crucial roles. The following quote from Lecturer A helps illustrate this point as far as students in the Translation and Interpretation majors are involved.

“Students in the English language program (English for Translation and Interpretation) are trained to translate materials from English to Vietnamese and vice versa. So, if they cannot understand the materials, it is impossible for them to do the translation.”

Not only was English reading perceived to be important in the academic programs, its relevance to the professional domains was also acknowledged as lecturer A went on: “ ... for their future jobs as well, if they want to become a professional translator, they need to be good

comprehenders at first. So, I think reading is a foundation skill upon which to build many other English skills.”

Similar points were reiterated by the student interviewees who have experience with the language demands at university and some initial experience in the respective workplace. Student A, for example, recounted her experience of the transition from high school to university and from university to the professional domain of teaching, during which English reading ability enabled her to address the emerging challenges.

“At high school, when talking about learning English, we talk about learning grammar, vocabulary, and some reading, no speaking, no writing, and very little listening. So, when I went to uni, grammar, some vocabulary and limited reading skills were what I had. These skills were the starting point for my English learning journey at uni. I started doing more prescribed reading exercises and searched for extra reading materials to read not only to improve reading skill, but also to find strategies for improving other skills. So, I reckon, reading was the first skill that I developed at uni, then it helped me improve other skills. And what’s more, I not only have to read to understand passages and to pass exams but also to teach reading skills to my students later on – it’s a different matter, much more difficult.”

While reading ability helped the student expand on her limited knowledge of the language to develop other English skills during transition from high school to university, it played another role in her transition from university to the workplace – to equip her with required knowledge and skills to teach her students how to read English materials. Without a good reading ability, she would have difficulty teaching reading skills to students.

Student B’s account seemed to support lecturer A’s comment on the importance of English reading ability in the English for Translation and Interpretation program.

“All four skills of reading, speaking, writing, and listening are important, but I think in the Bachelor program of English Translation and Interpretation, reading is particularly important. We have to do a lot of translation practice and comprehension of translation materials is a must. If I don’t have a good enough reading comprehension skill, I cannot do my job now.”

As indicated in the account above, the view that reading ability is a basis upon which the core ability in the program – the translation and interpretation ability – is built was underscored. A sound reading ability enables him to develop effective translation and interpretation skills. In contrast, if his reading ability is not strong enough, he will encounter challenges both in the academic programs and in his profession.

The ability to effectively read in English was also considered important for non-English major students because it was not only a requirement in their program but also in their future job as mentioned by one lecturer: “... for example, students in the Finance and Banking faculty need to achieve level 4 (Level B2) before they can graduate from university ... they need to be able to read and understand transaction emails, conditions in a business contract, or content of an invoice, to name but a few.” (Lecturer C). The importance of English reading ability for these students, however, was perceived to a much lesser extent as compared with their English major counterparts. This could be because English is considered a secondary subject in their curriculum with all the other subjects being offered in Vietnamese and what mattered more for these students was whether or not they could pass the English exams rather than the proficient use of English as laid out in the course objectives. This was evident from lecturer C’s comment:

“From my teaching experience, students are more concerned about passing exams than improving their English reading during the courses. And usually they are more concerned about their specialised subjects than about English ...”.

The perceived importance of the English reading ability in the academic programs as revealed via the interviews above foregrounds the amount and type of English reading required of students across different academic disciplines which are presented below.

9.2.2. The amount and type of English reading required in the academic programs

A corollary of the perceived importance of English reading by the participants was the amount and types of reading that students are required to do in their academic programs. In this respect, both lecturers and graduate students provided different accounts of the amount and type of reading required of students in the relevant academic programs.

For English major students, the message is quite clear: since English is their major, they are constantly exposed to a large amount of reading materials, not only in the reading skill module but also in other language-related modules. Specifically, English major students, regardless of their specializations, spend their first two years at university familiarizing themselves with the basic skills of reading via the prescribed materials as revealed via Lecturer B’s comment:

“According to the curriculum, the first two and a half years is focused on helping students become familiar with basic skills of reading. The amount of reading is quite balanced, and the topics and types of reading are neutral, using imported commercial learning materials so that they don’t need any specialized knowledge.”.

However, as students advance towards the end of their undergraduate programs, the amount and type of reading that they are required to do change in accordance with the requirements of the curriculum and the need to simulate the types of reading that they are likely to encounter in their respective future jobs. For example, English pedagogy students are required to read materials that equip them with necessary knowledge for their teaching career as commented by one Lecturer: *“Since the third year, they need to do more reading, in the reading subject and in other subjects as well. For example, they need to read British culture and teaching methodology, ...well both theories and practice of teaching..., materials required by their teachers and then do presentations in class. So basically, a lot of reading. (Lecturer B).”*. On the other hand, English translation students are prescribed reading tasks and materials relevant to their future job as a translator: *“... they are given technical materials to translate, such as topical news, business contracts, or travel itineraries. They not only need to comprehend the materials but also to translate them into Vietnamese so that lay people can understand”* (Lecturer C). In addition, a theme that emerged from students’ accounts was that their reading is not confined to prescribed materials only. They also find extra materials that enabled them to grapple with the demands of the curriculum. This is clarified by Student A’s comment.

“Yesterday when I reviewed learning materials to prepare for today interview, I was like startled because I don’t know how I could finish all those readings. This is not counting a large number of reading from external resources (primarily from the internet) that I did to understand the learning materials to be able to present in front of the class. I also did a large number of extra reading exercises and practice tests to improve my reading skill and to prepare for exams. So, among all skills, I spent most money on buying reading materials during my undergraduate program.”

In contrast to English major students, non-English major students were thought to engage in far less reading while types of reading were perceived to vary widely depending on specific disciplines. This is in part due to the curriculum requirements that prioritise other more discipline-specific subjects which are taught entirely in Vietnamese. Lecturer C, for example, attributes students’ lack of reading practice to the assigned curriculum.

“Although reading comprehension is important for them, I cannot expect them to do a lot of reading. You know, the curriculum only allows three to four hours of in-class English learning each week for students in disciplines other than English. The majority of class time is spent on grammar and vocabulary practices with very little reading comprehension. When they take the last English module, mostly in the first semester of

year three, they are given reading comprehension exercises but with very short reading passages. After that they part away with English at least to the extent that the curriculum is concerned.”.

When asked about whether students did extra reading to compensate for the lack of reading practice in class, he was rather sceptical: *“I also don’t think that they do extra reading because they have many other specialized subjects to take care of. Perhaps they only start to do intensive reading when they prepare for the exit exams.”.*

The students’ accounts seem to clarify the points made earlier by Lecturer C. Student C, for example, reported little engagement in reading practice inside and outside the classroom either due to the curriculum which put too much focus on grammar and vocabulary or to the perceived importance of speaking skills for his future jobs:

“What I can recall is the learning of grammar points and specialized vocabulary, most of the time. I did do reading exercises when I prepared for end-of-term exams and the exit exam, but not much. I don’t do any other reading besides what was required by teachers. I only practiced speaking English outside the classroom.”

In terms of text types, the reading practice of non-English major students also varied widely with respect to their academic disciplines. Lecturer A’s account below seems to indicate that as students approach the end of their undergraduate programs, they are required to read very different types of texts that reflect the reading they will encounter in their prospective careers.

“Reading types, hm, different disciplines will have different types of reading. For example, Finance and Banking students read short, concise transaction emails, letters, or contracts. Chemistry students read passages about chemical phenomena and reactions. Psychology students read passages about psychological concepts and practices. However, these types of specialised reading only come after they have achieved basic levels of English reading comprehension and are familiar with some of the specialized vocabulary in the field. Only students in the last reading module do these specialised reading.”.

It is indicated in the comment above that although graduate employability was taken into account in the curriculum design, the requirements that students be exposed to different types of reading relevant to their prospective careers came quite late in the program, raising concerns about the language resources needed to prepare students for the job markets.

In sum, via lecturers and graduate students’ interview responses, it was revealed that the perceived importance of English reading skill and the amount and types of reading required

in the academic programs vary with respect to academic disciplines. While students in the English major programs, both English teaching and English translation/interpretation, were required to read a large amount of English materials and were encouraged to read beyond prescribed materials, either to improve English reading skill per se or to supplement other language-related subjects, students in the Non-English majors are relatively restricted in the amount of reading and the types of reading required. A major factor that governs this fundamental discrepancy, as revealed from the interviews, was the assigned curriculum that positioned English either as a primary or a secondary subject in the respective disciplines. The lecturers, while acknowledging that English reading is an important component in the curricula for non-English major students, found themselves in a dilemma about how best to equip students with sufficient reading skills prior to their graduation within a limited time frame.

9.2.3. The reading tasks and skills required in the academic domains

This section reports findings related to lecturers and graduate students' perceptions and experience regarding the types of reading tasks that students commonly encountered in their formal study and the reading skills required of them in the academic domains. Two salient themes developed from the interview data, the variety of reading tasks included in the prescribed reading materials as well as in the end-of-term exams, and the different reading skills that students need to have in order to function properly in their respective academic programs. Each of these perspectives is discussed below.

A consensus among all interviewed participants was the variety of reading tasks encountered in the academic programs irrespective of their disciplines. The most common reading tasks were Short Answer Questions (SAQs) which require students to use information retrieved from reading materials to answer a given question and multiple-choice items (MCQs) which require students to choose the most appropriate answer from a set of given options. SAQs were so common that it has become a norm when it comes to reading comprehension, as perceived by all the interviewed lecturers and graduate students. Lecturer A and Student A accounts below help clarify the widespread use of SAQs in the English classes.

Lecturer A: "You'll see SAQs everywhere, in the textbooks, exercises, exams, and tests, in the reading curriculum per se or in any other subjects. When teachers want to assess students' comprehensions of assigned reading exercises or materials, they ask questions."

Student A: "Answering reading comprehension questions is most popular. It is in all the lessons, not only in reading lessons, but also in all other theory-related and skill

training subjects. For example, in a pronunciation training class, I need to read and understand the prescribed materials in advance. In class, the teacher checks our understanding by asking questions before the training takes place.”.

While SAQs are the most widely used reading task in almost all English classes, multiple-choice items (MCQs) is the most utilised task in the reading skill training module. Virtually all reading exercises and tests, as reported by the interviewees, have at least some sections that require students to choose the correct answer from a set of given options. In the excerpts below, Student B and Lecturer C explain the widespread use of MCQs in their learning and teaching practice as well as in standardized English proficiency tests that they have experience with.

Student B: “Yes, obviously in reading classes, MCQ is the most popular task. I was trained to answer a variety of reading tasks during the bachelor program, but MCQ received the most attention. In any of the reading tests I have taken before, I met MCQ in all of them. To be honest, I am pretty good at tackling this type of reading task because I spent most of my out-of-class time practicing this task.”.

Lecturer C: “In any reading tests whether at local or international levels, such as the IELTS, TOEIC, or TOEFL iBT, students will encounter this reading task. Having sufficient strategies to deal with this type of task will give students huge advantages. Therefore, during my teaching, I choose this task type and help students figure out ways to answer them. This is also to prepare students for future tests.”.

In addition to SAQs and MCQs, other reading tasks that emerged from informants’ responses include gap-filling, heading matching, summarizing, and true/false/not given, each of which has different variations depending on specific reading texts. A summarizing task, for example, may require students to write a short summary of a reading text or to complete a table. The majority of these tasks are commonly encountered in reading comprehension exercises and tests as part of a regular reading skill training class and in exam preparation classes, but less so in other English classes where students’ understanding of prescribed reading materials is usually assessed by SAQs or in-class presentations. Therefore, in order to help students grapple with these reading task types, teachers usually have to focus on training students to use appropriate reading skills and strategies: *“These tasks are very popular in the reading classes that I taught. They are included in the prescribed reading materials, but the strategies offered to deal with each task type are limited. So I have to find extra materials and exercises to help students learn the appropriate strategies to answer each of them.”* (Lecturer B).

A major concern that can be observed from the interviews with students was that the focus on the skills necessary to respond to these tasks was sometimes so intensive that students were more concerned with the task rather than comprehension. Thus, what was left after the classes were the techniques to deal with tasks rather than the text comprehension per se. This point was evident from Student B's account: *"Basically, I try to apply the strategies/techniques instructed by my teachers in the reading exercises in class, and then at home... Sometimes, I can answer correctly many questions, but after that I don't remember what I have just read."*. In addition, for the interviewed students, particularly those in the English major programs, a large amount of their out-of-class reading time was spent on practicing these reading tasks with a view to enhancing their performance in exams rather than on reading for pure comprehension and pleasure. This exam-oriented practice was evident from Student C's reflection of his learning experience:

"Looking back at what I learnt during the undergraduate programs, these task practice habits are good because they help me achieve high scores and pass exams. But if you ask if my reading experience at uni was enjoyable, I would say no. I prefer reading novels and short stories in English purely for relaxation as I am doing now to reading to pass exams."

Regarding the reading skills that students are required to develop to be able to function well in the academic programs, three main themes emerged from participants' interviews: reading for basic comprehension, understanding inferences, and reading to learn.

Reading for basic comprehension was considered the foundational skill of reading that students in all disciplines were expected to master in their programs. This skill encompasses their ability to understand specific and factual details explicitly stated in the texts, to understand simple ideas presented in the texts, and to gain a general understanding of what a text is about. Lecturer B provided more details about these basic reading skills as initial requirements for English major students in their first year at university:

"Well, understanding basic details – explicit information, facts and figures, simple details – these are compulsory requirements for student from their first year at uni. The reading syllabus for first year students are designed to help them at least achieve this... they come to uni with some experience of English reading from high schools and English is their major... if they don't have these basic abilities they will have a lot of difficulties later when they need to read a lot of materials in other subjects."

This requirement was also well appreciated by the students, particularly those in the English major programs. As revealed via Student A's comment below, the transition from high

school to university was difficult since students were not familiar with the reading skills as required in the university programs. It was this basic reading skill that students became first familiarized with, which enabled them to build up their reading skill repertoire.

“Understanding basic details is a must. I remember in my first few days at uni, I was asked by teachers to buy a lot of books to prepare for classes, all of them are in English. I was like shocked because at high schools I had never seen this much English... Well, the high school English exam had a reading comprehension section but I was taught entirely in Vietnamese to answer them by my high school teachers, you know word-by-word translation rather than reading comprehension. So when I held the reading textbook in my hand, I didn't know what to do, where to start. Over the first semester at uni, I gradually became more familiar with English and build up my initial skill in reading, that is reading for details. Most of the practices and exercises were focused on this skill” (Student A).

Depending on specific disciplines, the time at which students are expected to independently and proactively use this skill in reading comprehension may vary. Students in English major programs, as indicated in the quotes above, are expected to be able to use this skill effectively from earlier on in their academic programs while non-English major students are expected to spend most of their time at university developing this skill, as Lecturer C explained:

“That is what I expect non-English major students to be able to do. The English curriculum for them was also designed with this in mind ... it's good that by the end of their undergraduate program, they can understand details of a business letters, you know like numbers, figures, or main message, or to understand details in a business contract. For some disciplines, understanding complex details or complicated ideas is not necessary. For example, physical education students don't really need to use a lot of English in their future jobs. Primary education students also teach English to their future students, but that English is simple, just pronunciation, simple vocabulary, and basic greetings. Chemistry, physics, or mathematics education students only need the ability to read complex English if they decide to pursue higher degrees abroad, and so on...” (Lecturer C).

The differences in the perceived requirements of basic English reading skills and in the expected development trajectory of these basic skills between English and non-English major students may again ensue from the non-English curricula that place huge emphasis on the mastering of grammar, vocabulary, and simple structures of English and the limited time

students have for English learning in class: *“I cannot expect them to do more than that, you know they only have three hours a week for learning English in class.”* (Lecturer C).

Making inferences and reading to learn, as explained by the interviewed lecturers, are higher-order reading skills (Khalifa & Weir, 2009; Nassaji, 2003) that are perceived to be more important for English major students than to non-English major students, and thus an essential requirement for the former. Making inferences involve the ability to infer the meaning of a text or parts of a text based on information in the text and students’ prior knowledge, such as to figure out meaning of an unknown word, to recognise the implicit meaning of a detail or argument in the text, and to understand the author’s purposes, attitude, or opinions. Reading to learn goes a step further to use what is understood via the text for a particular purpose such as for presentation, for explaining, for applying in practice, or for making critical comments, depending on the context in which learning takes place. Of particular importance for this reading to learn skill is the ability to summarize and organize information in the reading materials to facilitate the follow-up activities. This also involves the ability to read selectively – to identify key information that is particularly relevant to the follow-up tasks for reading. Lecturer A’s comment below provides an example for the requirements of the above-mentioned reading skills for English major students

“Obviously, (English major) students need to be good at these inferencing skills, particularly in the reading classes... yes, there are a lot of exercises to help them improve these skills because they will meet them quite often in reading tests. The more they practice these skills in class the higher the chance that they will achieve high scores in reading tests... that is in reading classes. In other classes, such as teaching methodology, British and American culture, inferencing is less important because it is factual information and details that students need to understand... oh right... in these classes, students need to make presentations and to model teaching activities. So they need to understand the materials and use that understanding to help them do presentations or teaching activities.”

Although both inferencing and reading to learn skills are considered important for English major students, the complexity with which these two subskills are used vary with respect to the specific subject and reading types. Inferencing skill, for example, are required in both British literature classes and reading skill training classes, but the level of complexity with which this skills is used in the British literature class is different from that in the reading class primarily due to the different genres of reading typical of those classes. This is illustrated in the following excerpts by Lecturer B.

“In some classes such as British and American literature and reading skill training, inferencing is used quite often... but the kind of inferencing in literature classes is different from the kind of inferencing in reading skill training classes because the language in literature classes is mainly figurative. From my teaching experience, students have many difficulties in using inferencing skills in literature classes, and usually student learn by heart what I explained. Only a few students who have special knack for literature can do it.”

As for non-English major students, anything beyond the ability to comprehend basic information, as revealed by the lecturers, would be challenging for them. However, it was believed that the ability to read for follow-up tasks would be useful for their future jobs, as disclosed by one of the lecturers.

Lecturer C: “They would need some basic inferencing skills, but not much, primarily to do well on tests and exams. But I expect that their future job would need some skills like summarizing and organizing information. For example, students who will work in finance and banking would need to read English materials to write a report, or to read English reports by their employees to prepare for a meeting or conference... yet this is only possible if they work for international companies... but you know, if they decide to live and work in this (small) province, chances are low.”

In sum, the lecturers and graduate students’ accounts revealed some variations in reading tasks and reading skills required of students in the academic programs with respect to their specific disciplines. English major students are required to tackle different task types and to use a variety of reading skills in their academic programs. While the majority of reading tasks are encountered in the reading skill training classes and in tests and exams, students’ ability to employ different reading skills is important not only in reading tests but also in other language-related subjects. It is commonly assumed that in order to perform adequately in the academic domains, English major students need to have the ability to understand basic information, to make inferences, and to summarise and organize information for follow-up tasks. Non-English major students, though encountering similar range of task types, have lower expectations in terms of reading skills they need to achieve for their academic programs. Understanding basic information and simple inferencing skills are what students need to perform well in their academic programs, primarily to pass English reading tests while reading to learn skill, as expected by the lecturers, is more relevant to their respective future jobs.

9.2.4. Comparability between the academic domains and the L-VSTEP reading test

In order to establish the comparability of reading tasks and skills tested in the L-VSTEP reading test and those required in the academic domains, further exploration of the lecturers' and graduate students' perception and experience was conducted. To facilitate this process, all interviewees were given a sample L-VSTEP reading test and the list of reading skills assessed in the L-VSTEP reading test, as identified and examined in Chapter V and VI. They were then asked to comment on how relevant those reading skills and reading tasks are to the academic domains of the students. Analysis of the interview data suggested two salient themes, the limited task types presented in the test and the differences in terms of reading skills required in the academic domains and those found in the test, particularly when academic disciplines of the students are considered.

A general consensus among all interviewed participants was that the test method included in the L-VSTEP reading test was too limited as compared to the variety of task types that students are required to engage with in the academic programs. In fact, multiple choice is the only test method used in the L-VSTEP reading test. Participants were asked to further comment on how the differences in task types between the two domains might affect students' performance. In this regard, the three lecturers believed that students would not have much difficulty in terms of task type when they take the test because they are familiar with multiple-choice questions and spend much of their time in reading classes practicing this task type.

Lecturer B: "It's no big deal. Students are definitely familiar with this. You know, many of their reading exercises in the curriculum are of this type. I trained students to cope with this type quite often in the reading courses. Perhaps, they are more familiar with this type than any other tasks."

Lecturer C: "No problems for them (non-English students). They meet this type of task in their programs. Furthermore, they all, I believe, take test preparation classes before they take the test, so I think their teachers know how to help them deal with this task type."

The graduate students also agreed with this view. Student B, for example, claimed that: *"I don't have any problems. Everyone is familiar with this task, I think. Honestly speaking, as I know all the test items are of multiple-choice format, I spent most of my preparation time practicing this task type."*

However, when asked if the use of only one test method (MCQ) could elicit reliable information about students' comprehension of a text, all participants admitted reservations. The most common view was that the more task types were included, the more reliable the test would be since students who had good strategies to deal with MCQ would not be unfairly advantaged

over those who did not, and hence students' performance would be more fairly assessed. This is evident in Lecturer C's comment below.

Lecturer C: "I know there are students who take special private test preparation classes in which they are taught test-wise strategies, particularly how to deal with MCQ. The teachers are those who have high results in international proficiency tests such as IELTS or TOEFL. They just focus on honing students' test-wise skills... yes, they don't care about students' reading development. Just practice, practice, and practice. They attract students by advertising that students are guaranteed to get high scores on the test, and then force them to practice those skills. If there are more task types in the test, I think that students' performance will be less affected by test-wise strategies; or at least there should be other tasks that prevent test-wise practice. I feel frustrated with that."

It could be inferred from Lecturer C's account above that the use of only one test method not only compromised test fairness but also induced negative washback on the way that language learning and teaching was conducted. The biggest concern, therefore, was that students were trained to become "test-taking machine" rather than critical learners who were able to understand and evaluate reading texts themselves.

Lecturer B shared similar concerns about the use of only MCQs method in the test to assess students' comprehension as opposed to the variety of ways they can use in class to check students' comprehension: *"You know, in class, I have many ways to assess students' understanding of reading materials. In the test, there is only one way, and it's not sure how much of what students answer is actually their comprehension or just lucky guesses, or both."*

In addition, the use of different task types, as perceived by the students, might also reduce the potential threats of construct irrelevance variance as opposed to the "mechanical" nature of MCQ, thereby contributing to a more reliable and meaningful assessment of students' reading comprehension. When asked whether they took any commercial test-wise classes and whether they learnt anything from those classes besides test-taking techniques, Student B responded:

"I don't know about any test-wise classes, but I just think that there should be other task types that more accurately assess students' text comprehension. For me, if you ask me if I remember the reading topics in the test or recount what the reading texts are about, definitely no... because you know the test is only sixty minutes and what I do during the test is read and choose the answer, do it very quickly and anxiously, you know, like a machine."

In terms of reading skills in the academic domains and in the test, there were some disparities in participants' responses, particularly with regards to specific disciplines. The most notable difference was between English major students and Non-English major students in terms of reading skills required in their respective programs and in the test. English major students were believed to be exposed to a relatively similar skill profile in both domains. Lecturer C, for example, when looking at the names of the reading skills included in the test and their descriptions, simply responded that:

“The labels were different but from the descriptions, I see that they are similar to what students were taught in classes.”. Lecturer A, on the other hand, provided specific examples of the similar reading skills in the programs and in the test: “well, the skills are not new to students. All of them have been taught in the reading classes and also practiced in other subjects as well. For example, summarizing textual information and integrating information... they need to do this to prepare for class presentation in culture, literature, and methodology classes... yes, they need to find extra resources and synthesize information from these resources to prepare for their classes.”.

Not only were the skills perceived to be relevant to the academic domains but also they were thought to be well aligned with students' future jobs. Lecturer A went on: *“The same for their future jobs ... well I can see the coherence between what is taught, what is tested, and what is used afterwards here... If students become a tour guide or a teacher later, they still need to use these skills (Summarizing information and integrating information).”.*

This view was echoed by one of the graduate students. When asked if he still used the reading skills assessed in the test in his current professional work, Student B responded by giving an example of the use of integrating information skill.

“Yes, I do. Like in my current job, sometimes I have to translate a lot of unique materials into Vietnamese, there are concepts that I cannot find Vietnamese equivalents or I don't understand, then I need to google them, synthesize information from different sources to translate the text as closely in meaning as possible. It takes time, but that's how the job is done, honestly.”.

In addition, there was a concern expressed by one of the lecturers about the need to include a task that simulate the learning process in the academic programs. This concern developed from the inclusion of the two reading skills just mentioned above.

Lecturer A: “I think there needs to be more than that (the testing of summarizing and integrating information skills)... because you see in their programs, they do not just stop at summarizing or integrating information ... they do this to accomplish

something else, another activity, like presentation, model teaching, or writing assignments. This is more meaningful than just reading to show their ability to read.”.

It can be inferred from Lecturer A’s comment that the testing of isolated reading subskills (integrating and summarizing information subskills) as in the L-VSTEP reading test might not closely align to the target language use domain, a typical feature of which is the integrated nature of tasks that requires students to use different macro skills in tandem to accomplish the tasks at hand. From the perspective of the argument-based approach, the failure to simulate tasks in the target language use domain in the test seems to be a major concern that might undermine the dual-ground basis for inferences – the reconciliation of the task-based and competency-based perspectives for score interpretation and use (Chapelle et al., 2008).

For non-English major students, the higher order reading skills such as making inferences, understanding author’s purposes, summarizing and integrating information, were perceived to be beyond their grasp unless they took classes that focused on training and preparing them for the test.

Lecturer C: “I think there are only a few items that they (non-English major students) can answer. The majority of the items require them to use skills that are not included in the prescribed textbooks or practiced in classes. Like these items, “what does the word ‘she’ refer to ...”, “how much is the cost...”, I think they can answer them because these ones are basic...”.

Lecturer A: “I teach in the preparation classes for the L-VSTEP test, and when they first came to the class, they showed very limited range of reading skills, just basic reading. I had to work really hard to help them improve. Much more work for me in these classes than in my regular classes with English major students.”.

The non-English graduate student’s accounts provided further support for Lecturer C’s comments about the items that test skills beyond coverage in the curriculum. Student C, for example, when asked about the items that tested higher-order skills, responded that:

“Really, really hard for me. You know I felt relieved when I passed the test. The reading section was too difficult for me. Although I was taught these skills in the preparation classes, I didn’t know if I have applied them correctly in the test, I made many guesses.”.

In a nutshell, participants’ responses to the interview questions revealed that there were some misalignments between the test domain and the academic domains in terms of both task type and reading skills. Only one test method – MCQ - is used in the reading test as compared with the variety of reading tasks encountered in the academic programs. This important

difference notwithstanding, both lecturers and students in the interviews agreed that students achieved some familiarity with this task type during their programs through skill training and classroom practice. Reading skills, on the other hand, were perceived to present more problems, particularly for non-English major students since their academic programs provided limited coverage of the skill range encountered in the test. The problems were less intense for English major students as they were believed to be familiar with most of the tested reading skills.

9.3. Discussion

The purpose of this chapter was to provide additional evidence to address the extrapolation inference in the argument-based framework which makes an assumption about the congruence between students' performance on the test and their performance in the target language use domains. While the previous chapter looks at the test score patterns in both domains, the findings reported in this chapter examine the correspondence in task types and reading skills between the two domains. To this end, semi-structured interviews were conducted with three lecturers and three graduate students who had experience with both domains – the academic programs and the L-VSTEP reading test.

Analysis of the interview data suggested that English reading ability was perceived to be an important skill, particularly for English major students, not only in the academic programs but also in the future professional domains, thus supporting the need to assess their reading ability before graduation. The amount and type of reading in the academic domains were believed to vary with respect to specific disciplines. English major students were expected to engage in a large amount of reading in their bachelor programs while non-English major students had lower expectations, primarily due to the curriculum constraints and the role of English in the relevant bachelor programs.

In terms of reading tasks, a notable discrepancy was observed between the test and the academic domains. Multiple-choice questions based on a common reading prompt is the only test method used in the test while the academic programs included a variety of tasks that students were required to perform. Another prominent difference was the range of reading skills assessed in the test and those found in the curriculum for non-English major students. While the academic programs for non-English major students were focused on training basic skills of reading, the test included a wider range of reading skills, the majority of which are at higher level of reading processing. On the other hand, the reading skills assessed in the test were considered to be well aligned to the curriculum for English major students, though the inclusion of the reading to learn task was believed to provide a more accurate assessment of

students' reading ability. The above-mentioned incongruences in terms of reading tasks and reading skills in the two domains could be explained by two salient factors, the context of reading and the discipline-specific requirements, as revealed via the participants' accounts.

As for English major students, the academic context in which reading occurs has some unique features that set it apart from the test context. In the academic programs, students need to perform certain reading tasks which require particular reading skills not only to improve reading per se but also to advance their knowledge in specific subjects. Thus, to some extent, reading comprehension is considered a means to an end rather than an end in itself. As explained by the lecturers, English major students are required to do presentations, model teaching or write assignments in the British/American culture and teaching methodology classes, and reading skills constitute an important component in the process as they need to search for, select, integrate, synthesize, and organize information from different reading resources. All these reading skills are conducted in concert with a particular purpose in mind rather than an arbitrary assemblage of certain skills to answer a particular reading comprehension question.

In addition, students are encouraged to critically evaluate the reading materials as well as to distinguish important pieces of information from less important ones for selective reading. On the other hand, in the test, the focus of assessment is predominantly on the concrete reading skills and students' comprehension and interpretation of texts are constrained by predetermined response options. Therefore, the ability to evaluate and integrate multiple reading sources for critical reading and follow-up tasks seems to be missing in the test. This point is evident from Lecturer A's and Student B's comments about the involvement of reading skill in the accomplishment of tasks in other language-related subjects such as British culture and Teaching methodology. Similar phenomena were also observed by Barton (1994) and later by Alderson (2000) who argued that reading in the academic studies was rarely a stand-alone process without any relation to other academic activities (p.148). Yet, the question of how such real-world authentic tasks could be constructed in tests remains to be explored.

Differences in discipline-specific requirements for English also play a large part in the reading tasks and skills disparity between the two domains. What was consistently noted during the interviews was the clear differences between the English major and non-English major students in terms of the perceived amount and type of reading required of them in the academic programs, the extent of exposure to reading materials in their programs, and the reading skills instructed in the relevant programs, all of which were somewhat governed by the assigned English curricula. While English is the primary medium of instruction and the focus of practice

both in and outside the class for English major students, their non-English counterparts only have three contact hours of English per week, which normally spreads over a four-semester period. This amount was believed to be far from sufficient to equip students with necessary skills and knowledge for the test unless supplementary test preparation classes were undertaken. These disadvantages, according to the interviewed participants, manifest themselves in the limited experience with and exposure to the reading tasks and skills commonly encountered in the test. This aspect of the curriculum inadvertently put non-English major students in a disadvantaged position when it comes to test performance as compared to their English major counterparts, raising concerns about the fairness of the test for this particular cohort of students.

Another noteworthy factor that might have contributed to this English and non-English discrepancy was the socio-economic context in which English learning takes place. The institution where the current project was conducted is situated in a small province in central Vietnam. There are not many international corporations or companies in the area that require the use of English for graduates in disciplines other than English. This might have impacted students' perceptions of the importance of English and made it a peripheral priority in the employability skill set of non-English major students. As commented by one of the lecturers, unless (non-English major) students plan to find a job in a metropolitan area, the ability to use efficiently a variety of English reading skills, particularly those at higher-order levels, is considered non-essential. This relatively peripheral role of English is embodied in the curricula for non-English major students where all the discipline-specific subjects are delivered in Vietnamese.

In summary, this chapter provides insights into relevant stakeholders' (i.g. teachers and students) perceptions about the alignment between the reading tasks and skills assessed in the L-VSTEP reading test and those required in the relevant academic programs. In this respect, the English reading curriculum for English major students seems to be relatively well-represented in the test while non-English major students may find the test challenging unless they take intensive test preparatory courses. This offers useful implications for teachers, students, curriculum designers, and policy makers at the relevant institution, which is discussed in more depth in Chapter XI.

10.1. Introduction

This chapter articulates the validity argument for the L-VSTEP reading test. As discussed in section 3.2, the validity argument provides an evaluation of the coherence, completeness, and plausibility of the warrants, assumptions, and backings as laid out in the interpretive argument by drawing on the variety of empirical findings that have been reported so far. More specifically, since the primary focus of the research project is on the explanation and extrapolation inferences, the validity argument presented in this chapter offers an evaluation of three assumptions that support the warrant for the explanation inference and two assumptions that support the warrant for the extrapolation inference, drawing on empirical results either as backings for rebuttals. The former includes evidence that support the inferences while the latter encompasses evidence that weakens the strength of the inferences. Each of the warrants, assumptions, and relevant backings or rebuttals for the two inferences are summarized in Table 10.1 and are discussed in detail in the sections that follow.

Table 10. 1. The validity argument for the L-VSTEP reading test

Inferences	Warrants	Assumptions	Backings	Rebuttals
<i>Explanation</i>	Students' scores on the L-VSTEP reading test can be attributed to the construct of general English reading proficiency	<p>1. Observable task characteristics underlie task performance consistency.</p> <p>2. Reading processes and strategies engaged by test-takers vary according to theoretical expectations.</p> <p>3. The underlying structure of test reflects highly intercorrelated components explaining theoretical expectations.</p>	<p>1. Text concreteness predicted item difficulty in the test.</p> <p>2. Test-takers' reading processes vary according to sub-skills tested, corresponding to the expert judgment on the skills and processes elicited.</p> <p>- Test-takers' reading processes vary according to reading proficiency levels.</p> <p>3. The underlying structure of the test was identified and well aligned with the guidelines for test development and expert judgment.</p>	<p>1. Item length and plausible distractors accounted for a significant amount of variance of item difficulty.</p> <p>2. Test-taking strategies, construct irrelevant items, and local dependence items were identified.</p> <p>3. The underlying structure of the test was found to be non-invariant at the configural level across participant groups with different proficiency levels.</p> <p>- The underlying structure of the test was found to be non-invariant at the</p>

				metric level across participant groups with different academic disciplines.
<i>Extrapolation</i>	The observed performance of students on the L-VSTEP reading test predicts their English reading performance in the academic programs at the institution.	<p>1. Students' test scores predicts their performance in the academic programs as assessed by their self-reported English reading proficiency.</p> <p>2. The reading tasks and skills as assessed in the L-VSTEP reading test are compatible with those required in the relevant academic programs.</p>	<p>1. A moderately strong regression coefficient from the latent construct of students' test scores to the latent construct of self-assessment was found.</p> <p>2. All reading skills as assessed in the L-VSTEP reading test were found to be comparable with those required in the academic programs for English major students.</p>	<p>1. Students' scores on the L-VSTEP reading test only accounted for a modest amount of variance in their self-assessment responses.</p> <p>2. Only one test method was included in the L-VSTEP reading test as compared with a variety of task types required in the academic programs.</p> <ul style="list-style-type: none"> - Reading skills relevant to the academic domain such as critical reading and synthesizing from different readings were not represented in the test. - The higher-level reading skills as assessed in the L-VSTEP were perceived to be incompatible with, and thus not reflecting those required in

the academic programs for non-English major students.

- The reading demands of the English major and non-English major courses (Vietnamese for non-English majors and English for English majors) were so fundamentally different that test fairness might be compromised.

- The likelihood of non-English major students actually requiring English reading skills in their jobs was low, if they stayed within the province.

10.2. The explanation inference

The explanation inference is based on the warrant that expected scores of students on the L-VSTEP reading test are attributed to the construct of English reading proficiency as specified in the guidelines for test development currently in use at the institution where the test is administered. This warrant is based on three interrelated assumptions relevant to the three-plane hierarchical model of abstractness for the explanation inference proposed by Chapelle et al. (2008).

Assumption 1

At the most concrete level is the assumption that observable task characteristics are identifiable and systematically influence task difficulty as theoretically expected. Backing for this assumption was established through the examination of the linguistic and discourse characteristics of the reading texts, items and item-text interaction, and how these characteristics influenced item difficulty of the tests. Results relating to this assumption were reported in chapter VII. Overall, among the 20 variables identified via extensive literature review, the Coh-matrix software for the automatic analysis of linguistic features, and expert judgment, only four variables were found to be significantly related to item difficulty of the tests, one text variable (text concreteness), one item-text variable (plausible distractors), and two item variables (lexical overlap between the correct answer and the alternatives and item length).

Except for the item variable of lexical overlap, the other three variables significantly predicted item difficulty as revealed via the multiple regression analysis. Plausible distractor was the strongest predictor of item difficulty, followed by item length and text concreteness. While the identification of text concreteness as a significant predictor of item difficulty provided evidence in support of the explanation inference, the finding that the item and item-text variables of lexical overlap and plausible distractor constituted the more robust predictors of and accounted for more significant amount of variance in item difficulty generated rebuttals against the explanation inference. The former suggested that item difficulty varied as a function of the linguistic features of the reading texts, at least to the extent that the concreteness level of the texts was concerned. The latter, on the other hand, implied that the item difficulty of the tests was unduly influenced by the construct-irrelevance factors associated with the test questions. In other words, the probability of a test item being accurately answered varied according to the length of the item stem and options, and with the number of distractors that could be directly confirmed in the texts, rather than with the linguistic complexity of the texts.

This variation in item difficulty was only minimally accounted for by the text variable of text concreteness.

All in all, it could be assumed that the explanation inference based on the assumption at the most concrete level of surface features of task characteristics was only partially supported with stronger evidence in the form of rebuttals. This finding seems to be in line with Chapelle et al. (2008) who, due to a lack of backing evidence, excluded this assumption from the interpretative argument on which the validity argument for the Test of English as A Foreign Language was made. This exclusion, according to them, did not undermine the explanation inference given the multiple levels of potential explanation for test scores. Rather it underscored the need to collect more evidence as backings for the assumptions at the more abstract levels of the explanation inference.

Assumption 2

At the middle plane in the hierarchical model of abstractness for the explanation inference is the assumption that the reading processes and strategies engaged by students during task performance varied with respect to theoretical expectations. Backing for this assumption was explored by the examination of the consistency between students' verbal report of the reading processes and strategies during test taking and those judged to be required in answering test items through expert judgment. Further exploration of the extent to which reading processes and strategies as reported by students varied in keeping with their reading proficiency levels provided additional backing for this assumption. Results of this phase were reported in chapter V. One salient finding was the general agreement among students and experts in terms of the primary reading subskills/processes assessed by the test items. The primary reading subskills assessed by the items as reported by both students and experts included understanding information explicitly stated in the texts, understanding cohesion based on connective devices, inferring word meaning, inferring situational meaning, integrating information across sentences and paragraphs, understanding author's attitude, purpose, and opinion, summarizing information, and recognizing organizational structure of the texts. These subskills were identifiable at individual item level from both expert judgment and students' protocols and covered the range of reading subskills stated in the guidelines for test development. This finding, therefore, provided evidence that supported the assumption that the processes, skills, and strategies students engaged in during task performance varied according to theoretical expectations.

Different perspectives among students' verbal protocols and between students' protocols and experts' judgment occurred when the potential involvement of multiple

subskills/processes in answering a particular test item was considered. These differences became more pronounced as students moved from items that required lower-level subskills/processes to items that required higher-level subskills/processes, and between high-achieving students and low achieving students. Items that primarily assessed higher-level reading subskills/processes (e.g. integrating information, summarising information, and inferring information) and that required the use of multiple subskills/processes during test taking did actually elicit a range of different subskills/processes from students as expected by experts. However, the skills reported by students when responding to these items varied from student to student and did not exactly match the skill profile expected by experts. In addition, the skill profile employed by high-achieving students to answer a particular item was different from that employed by low-achieving students. The former applied a range of skills similar to those expected by experts while the latter were rather limited and inefficient in the range of skills they used, with a heavy reliance on lower-level processes. These findings, though demonstrating some inconsistency between students' actual performance and experts' judgment in terms of reading skills/processes, did not necessarily weaken the explanation of the test scores given that students' utilization of these skills/processes varies with respect to theoretical expectations. To clarify, reading is a multi-component and multi-process construct that predisposes students to use a variety of processes, skills, and strategies in reading comprehension and reading test performance depending on specific task characteristics and learner variables (Cohen & Upton, 2007; Rupp et al., 2006). For example, reading comprehension tests with multiple choice item response format have a tendency to induce various response processes among learners in ways that may deviate from response processes in non-test contexts (Rupp et al., 2006). Variation among students in this study in the use of reading skills/processes to answer different reading comprehension questions was, therefore, not unexpected. As a result, these findings can be considered as backings that largely support the explanation inference.

Another finding that had relevance to the explanation inference was the noticeable pattern of test-taking strategies reported by students, particularly the use of the strategy of eliminating implausible answers. As discussed in chapter V, this practice might have ensued from the way the question was formulated (e.g. what is *not true* according to the passage, what is *not mentioned* in the passage), from the need to reduce the number of options down to a more manageable situation, or from a complete lack of comprehension. Neither of the aforementioned scenarios contributed useful information to the explanation of the test scores since they suggested that students approached the reading texts as a problem-solving task rather

than a comprehension task (Rupp et al., 2006). This finding coupled with the quantitative finding in chapter VII that plausible distractor was the strongest predictor of item difficulty seem to lend further empirical evidence to rebutting the explanation inference.

Analysis of the verbal reports also revealed several problematic items that manifested themselves to be potential rebuttals to the explanation inference. These include Item 4 which was subject to construct-irrelevance factors and Items 21, 22, 24, and 25 which exhibited local item dependence (for detailed discussion of these items, see chapter V). These items, therefore, stand out as potential candidates for item revision.

Assumption 3

At the most abstract level of explanation is the assumption that the underlying structure of the L-VSTEP reading test reflects highly intercorrelated components that explain theoretical expectations. Backings for this assumption were sought by the study that examined the factor structure of the test and the extent to which this factor structure reflected the theoretical construct of English reading as laid out in the guidelines for test development and with respect to the relevant literature in the field. Examination of the factorial invariance of the test factor structure was also undertaken to provide additional backings for the explanation inference. Results of this study were reported in chapter VI. Generally, the theoretically informed and empirically derived model of the test constructs was identifiable from 544 students' score patterns. More specifically, confirmatory factor analysis of parcel-level data achieved exceptionally good model fit with statistically acceptable and substantively meaningful parameter estimates. This model represented the underlying structure of the test with a general English reading proficiency factor accounting for variances in seven observed indicators of reading subskills derived from the guidelines for test development and expert judgment (see assumption 2). This finding provided evidence that students' test scores did reflect the theoretically informed structure of the test and therefore can serve as backing for the explanation inference.

Contrary to the factor structure of the test, the measurement invariance analysis of the test structure across different participant groups revealed evidence that might rebut the explanation inference. The factor structure of the test was found to be non-invariant across high- and low-achieving students, evidenced by the extremely poor model fit at the configural level. Further analysis at the group level revealed that the factor structure of the test was reproduced for the high-achieving group but not for the low-achieving group. As discussed in chapter VI, this finding could be attributed to a number of factors, such as low-achieving students' limited skill range and their overreliance on lower-level subskills as well as the group

homogeneity as a result of their poor performance on the test. The former explanation seems to gain further empirical support from the finding in the stimulated verbal recall (see chapter V and assumption 2 above) that low-achieving students reported relying on a limited skill range, inefficient skill use, and a tendency to draw primarily on low-level subskills to answer the test questions. Although this finding was interpreted as evidence backing the explanation inference in assumption 2, the statistical evidence retrieved to evaluate assumption 3 suggested that the test measured different subskills across student groups with different reading abilities, hence rebutting the explanation inference at the most abstract level of evidence.

The factor structure of the test was found to be invariant at the configural level, but non-invariant at the metric level across three groups with different academic disciplines, English pedagogy group, English translation group, and non-English major group. More stringent analysis at the metric level revealed two non-invariant factor loadings, those of understanding pragmatic meaning and summarizing textual information onto the general construct of English reading proficiency. A potential explanation for this finding was students' exposure to different reading materials in different academic disciplines. This was later confirmed via the lecturers' and students' interviews in the study that examined the comparability between the reading tasks and skills required in the academic domains and assessed in the test in an attempt to address the extrapolation inference (see below). It can be assumed, therefore, that this line of evidence only partially supports the explanation inference, with some evidence that weakens the strength of the factor structure of the test.

10.3. The extrapolation inference

The extrapolation inference is premised on the warrant that students' scores on the test reflect their English reading proficiency in the academic programs at the same institution where they are pursuing their undergraduate degrees. This warrant is built upon two assumptions, students' scores on the test predict their English reading proficiency in the academic domains and the reading tasks and skills required in the academic domains are comparable to those assessed in the test. Backings for these assumptions were addressed by research question 4 and 5. The following sections provide an evaluation of the evidence for the two assumptions that underlie the extrapolation inference.

Assumption 1

This assumption addresses the extent to which students' English reading proficiency in the academic programs can be predicted by their scores on the L-VSTEP reading test. Backings for this assumption were sought by the study that examined the structural relationship between

students' test scores and their self-assessed English reading proficiency in the target domain via a self-assessment questionnaire. This structural relationship denoted the hypothesis that students' performance on the test predicts their performance in the target language use domains, herein the academic programs that they were pursuing at the institution. Confirmatory factor analysis of the reading test structure model and the reading self-assessment structure model yielded relatively comparable model structures with adequate measurement properties that represented both types of data. This offered initial evidence that supports the extrapolation inference because the underlying structures of the reading test and the self-assessment questionnaire were found to be relatively similar, thereby enabling the comparability between the two domains. The subsequent structural equation model in which test data were regressed onto self-assessment data yielded acceptable model fit with appropriate parameter estimates, hence strengthening the argument made earlier regarding the comparability between the two domains. Of note was the statistically significant and moderately strong regression coefficient from test data to self-assessment data, indicating that the use of test scores to predict students' performance in the relevant academic programs was warranted. These findings can be considered backings for the extrapolation inference of the test.

One potential rebuttal to the extrapolation inference as retrieved via the study was the relatively large standardized residual value as a result of regressing test score data onto self-assessment data, suggesting that a large amount of variance in self-assessment (.88) was unexplained. As discussed in chapter VIII, potential explanations for this finding were the multitude of experiential and individual characteristics that might affect the accuracy of self-assessment. For example, final-year students in the current study might be tempted to over-evaluate their reading proficiency in an attempt to present themselves in socially desirable ways because the curricula expected students at their stage to be able to master certain reading skills. Lack of training before self-assessment might be another reason contributing to the observed phenomenon. The finding of large residual variance in the study was comparable to that found in Fan and Yan (2017) who also testified the legitimacy of self-assessment instruments at the general language ability level and flagged the need to probe further into ways to leverage the accuracy of self-assessment to reduce large residual variance in the structural equation model. In sum, findings in the study generally support the warrant that students' performance on the L-VSTEP reading test reflects their reading proficiency in the target language use domains. Some rebutting evidence was also identified, which suggested ways to enhance the accuracy of self-assessment practice to better represent performance in the target domains.

Assumption 2

Assumption 2 for the extrapolation inference is that the reading tasks and skills assessed in the L-VSTEP reading test are comparable to those required in the academic programs that the students are pursuing. Backings for this assumption were established via semi-structured interviews with both lecturers and graduate students who have experience with both the academic programs and the test. The primary focus of the interviews was on the participants' perception and experience about the comparability between the reading tasks and skills in the test and in the academic programs and whether the test was suitable for the target students. Results of this study were reported in chapter IX. Generally, participants perceived the test to be suitable for the students, particularly for English major students. Non-English major students, on the other hand, were believed to experience challenges unless they took intensive preparatory courses before sitting the test.

In terms of the reading tasks, lecturers and students perceived an underrepresentation of reading tasks that are encountered in academic programs in the test. Students in the academic programs encounter a variety of reading tasks, ranging from questions and answers, gap-filling, cloze test, to matching headings, true/false/not given, and multiple-choice items. On the other hand, the reading test only includes one task type – multiple-choice questions, raising concerns about students focusing excessively on training test-taking strategy rather than on improving reading skills. Additional findings from the studies based on assumptions 1 and 2 for the explanation inference further elucidated this concern. Accordingly, students were found to have a tendency to use the strategy “eliminating implausible answers” to answer the multiple-choice questions in the test (see chapter V and assumption 2/explanation inference); and the strongest predictor of item difficulty in the test was the number of distractors in the multiple-choice questions that could be directly confirmed or disconfirmed via the reading texts (see chapter VII and assumption 1/explanation inference). Therefore, test method presents itself to be a potential rebuttal to the extrapolation inference.

Reading skills as assessed in the test were perceived to cover the range of skills required in the academic programs for English major students. These skills were identified in the curricula for English major students not only in the reading skill training classes but also for other language-related modules. Moreover, the interviewed lecturers suggested another important skill in the academic programs that was not sampled in the reading test. That was the students' ability to use what they retrieved from the reading materials to perform follow-up tasks, such as making a presentation, modelling teaching techniques, or writing opinion essays. This integrated language assessment task was conceptualized as one of the reading purposes – reading to learn – for the reading construct in the development framework for the revised

TOEFL test (Cohen & Upton, 2006), thereby calling for more consideration for the inclusion of this task in the L-VSTEP reading test to better represent academic language use domains.

In contrast to English major students, non-English major students were believed to experience difficulties in the test since the range of reading skills assessed by the test were far beyond those required in their academic programs. Only a limited range of reading skills at lower levels were required of students in non-English majors, largely due to the constrained English curricula and the perceived role of English reading skill in the academic programs and in future jobs. This finding seems to lend more empirical evidence to explain the findings from the study based on Assumption 3 for the explanation inference that the factor structure of the test was non-invariant at the configural level across students with different reading performance levels and at the metric level for students with different academic disciplines. Non-English students were found to achieve the lowest mean scores on the test (see chapter VI) and the two non-invariant loadings across different academic disciplines were related to the understanding pragmatic meaning and summarizing textual information subskills – the two higher-level reading subskills.

All in all, findings from the study only partially support the assumption that the reading tasks and skills assessed in the L-VSTEP reading test simulate those required in the relevant academic programs. Evidence that potentially defeats the extrapolation inference entails the underrepresentation of test tasks and reading skills in the test, particularly for Non-English major students.

Based on the evaluation of the interpretive argument for the L-VSTEP reading test as presented in the validity argument above, the next chapter, chapter XI provides some theoretical, methodological, and practical implications for relevant stakeholders of the test including policy makers, curriculum designers, researchers, teachers, students, and test designers.

11.1. Summary of the research findings

Informed by the argument-based approach to language test validation (Chapelle et al., 2008; Kane, 2013), this study has articulated an interpretation and use argument for the L-VSTEP reading test and provided a general evaluation of the interpretation and use argument, drawing on multiple lines of empirical evidence throughout the project. General support for the explanation inference ensued from the evidence that students reported a relatively similar profile of skills use to what experts expected, that the factor structure of the test was identifiable as theoretically expected, and that textual features predicted item difficulty of the test to a certain degree. In contrast, the explanation inference was weakened by several rebuttals such as students' tendency to use test-taking strategies, non-invariance of the factor structure of the test across different student groups, and the predictor power of construct-irrelevant factors related to item features when it comes to item difficulty. Backings for the extrapolation inference include the statistically acceptable structural equation model that captures the predictive relationship between students test scores and their performance in the target language domains as assessed by their self-reported English reading proficiency, and the relatively comparable reading skill profile as assessed in the test and required in the relevant academic programs, especially for English major students. On the other hand, the explanation inference was rebutted by the relatively large amount of residual variance in the students' self-assessment and by the misalignment between the reading tasks and reading skills assessed by the test and those found in the academic domains for Non-English major students.

11.2. Theoretical implications

The current project has effectively deployed the argument-based framework for test validation to examine the validity of the interpretation and use of a locally-developed test of English reading proficiency, thereby extending the line of research that uses this framework in language testing and assessment and providing an example of how it can be utilized at the local level. As argued by Kane (2013), the validation efforts should give priority to the inferences in the argument-based framework that are questionable and that need more empirical evidence to provide a more informed interpretation and use of the test scores. The current project focuses on two inferences, the explanation inference and the extrapolation inference because the L-

VSTEP test is a newly developed test of English proficiency in a local context in Vietnam and there has been little empirical evidence regarding its validity published in the literature. The exploration of these two inferences in the argument-based framework is timely given that empirical evidence is needed to shed light on the extent to which the test measures what it is designed to measure and whether students' performance on the test can account for their performance in the relevant target language use domains. Therefore, the argument-based framework can be used to construct a validity argument not only for well-known high-stakes international English proficiency tests (e.g. TOEFL, IELTS, TOEIC) which have been extensively reported in previous studies, but also for English proficiency tests that are developed at local levels (Johnson & Riazi, 2017; Li, 2015a).

Since its introduction to the field of language testing and assessment, this framework has gained increasing popularity and informed a number of language test validation projects, the majority of which focus on tests of language proficiency that include all four macro language skills of speaking, listening, writing, and reading (see chapter IV). Fewer attempts are made at macro-skill level to construct a validity argument based on this framework, with a few exceptions such as writing (Johnson & Riazi, 2015; Mendoza & Knoch, 2018; Yan & Staples, 2020), speaking (LaFlair & Staples, 2017; Yang, 2016), and listening (Aryadoust, 2013; Pardo-Ballester, 2010). The current project can be considered among a few studies that construct a validity argument for an English reading proficiency test. Therefore, the interpretation and use arguments and the assumptions that underlie the two important inferences of explanation and extrapolation as articulated in this study may provide useful information for other similar language test validation, particularly those at local levels.

11.3. Methodological implications

A major consensus established in the general literature and maintained throughout the project is the multi-component, multi-process nature of L2 reading comprehension. The use of a single method to examine L2 reading, therefore, might not be able to provide a nuanced understanding of the multifaceted nature of this construct. In addition, situated within a paradigm shift in language testing and assessment from a strong focus on quantitative methods to a more balanced and flexible mixed method paradigm (Jang et al., 2014), the argument-based approach to test validation entails a chain of inferences leading from students' observed performance on the test to the decisions made on the test scores, each was built upon a number of assumptions that required multiple lines of evidence as backings. The two conditions mentioned above provide strong rationales for the use of a mixed method paradigm in the

current project to address the research problems. Indeed, the multiple lines of mutually supplementary empirical evidence, both qualitatively and quantitatively, generated by the research program in the current project have enabled an informed and thorough evaluation of the interpretation and use argument, thereby providing a more informative and meaningful interpretation and use of the test scores for the relevant stakeholders (see the following section). For example, the qualitative interview data with lecturers and graduate students revealed that the reading skills required in the English curricula for non-English major students were limited as compared to the range of reading skills assessed in the test. This evidence could be taken to shed light on the quantitative finding that the factor loadings of the two higher-level reading subskills of understanding pragmatic meaning and summarizing textual information were non-invariant across students with different academic disciplines. Another salient example was the concern raised by lecturers in the interviews that students may focus too much on test-taking strategies in response to the multiple-choice items in the test. This concern found resonance in the quantitative finding that the number of distractors in a multiple-choice question that could be directly confirmed in the texts was the strongest predictor of test item difficulty. Similar finding was also found in the stimulated recall interviews that eliminating implausible answers was a common strategy used by students across different levels of performance.

As important as the use of multiple methods were the collection of multiple data types from different stakeholders of the test and the combination of the two inferences of explanation and extrapolation. This practice has enabled a more comprehensive account of the interpretation and use of the test scores, with a balanced focus on both the technical issues of the test scores as informed by the explanation inference and on the meaningfulness of the test scores as informed by the extrapolation inference. The accumulation of mutually supplementary rather than self-explanatory evidence in the current project, therefore, provides a more robust background for the evaluation of the interpretation and use argument for the test.

11.4. Practical implications

An evaluation of the evidence derived from the current project, especially evidence that undermines the interpretation and use argument offers some potential implications for relevant stakeholders of the test including students, teachers, researchers, test designers, policy makers, and curriculum designers. These practical implications revolve around the central concept of test fairness and its associated components.

Test fairness has long been a major concern in the development and validation of language tests (McNamara & Ryan, 2011). Especially, the concept of fairness has recently been

promoted to a foundational position in testing and constitutes one of the most fundamental changes in the latest version of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014). There are many definitions of fairness, but the most useful one, as stated in the Standards, is “the extent to which the inferences made on the basis of test scores are valid for different groups of test takers” (p.19). This definition has implications for different aspects of the test development and validation process, two of which are informed by relevant empirical evidence in the current project and are discussed in the following sections, namely fairness in test design and fairness in test score interpretation.

A major concern in the test design phase is the anticipation of potential sources of score variance, of which construct-irrelevant sources are detrimental to test fairness. Consideration of construct-irrelevant sources are premised on the assumption that score variance should be primarily accounted for by factors that are relevant to the constructs under investigation, and therefore any factors that are irrelevant to the core constructs should be anticipated and avoided in the design phase. In consideration of the findings in the current project, the concept of test fairness seems to be compromised by several lines of evidence pointing to construct-irrelevant factors.

First, stimulated recall data indicated that students had a tendency to use the test-taking strategy of eliminating implausible answers. This suggested that the test has been constructed in a way that potentially induced students to resort to response processes irrelevant to the intended response processes. This finding seemed to be corroborated later by the quantitative finding that the most robust predictor of item difficulty was the number of distractors in an item that could be directly confirmed in the texts. Second, the detection of potential local item dependence in the stimulated verbal recall and of item length as a significant predictor of item difficulty suggested that the probability of a student answering an item correctly, to some extent, depended respectively on the answers to other items that inadvertently carried the required information and on the way the questions were formulated, rather than on the student’s ability under investigation.

The identification of these construct-irrelevant sources offers implications not only for test designers but also for teacher, students, and researchers. As for test designer, caution must be exercised when crafting test items so that construct-irrelevant sources are controlled for. For example, the language and length of the multiple-choice questions should be appropriately sampled so that they do not cause excessive difficulties for test-takers. Attention should be paid

to the careful selection of vocabulary, grammatical structures and sentence formulation that are commensurate with the proficiency levels of the target test-takers and that do not impose unduly cognitive load on test-takers' processing of the questions. In addition, careful scrutiny of the item content should be conducted as a post hoc procedure to identify any potential local dependence items. Priority should be given to items that are constructed on the same prompt so that any items whose answers give away the answers to other items can be detected and handled early in the test development process. As for researchers, consistent and transparent research attempts should be made on a continuous basis so as to gain more insight into aspects of the tests that are questionable and the various factors that might introduce construct-irrelevant variance. The argument-based approach as used in the current project can be a viable framework that enables a comprehensive inquiry into these aspects of the test. For example, the use of think-aloud protocols in combination with item response theory can aid in the detection of local dependence items. More sophisticated methods for item difficulty modelling, such as multilevel modelling, can be used to explore other aspects of multiple-choice questions that might introduce construct-irrelevant variance. More importantly, studies that look into the effects of test methods on test-takers' performance and the various task types that may elicit samples of test-takers' performance in a more reliable and valid ways are needed so that the types of task used in the test can better represent those required in the target language use domains. This may help strengthen the extrapolation inference for the interpretation and use of the test scores. As for teachers and students, the teaching and learning of reading comprehension should be focused more on comprehension per se and the ultimate purpose of skill training should be to develop students who have good reading ability, flexible skills use, familiarity with reading materials of different types and genres, and are able to apply what they read to accomplish a specific follow-up tasks in their academic programs. The practice of teaching and learning to the test should be reduced wherever possible.

A major concern in the interpretation and use of the test scores is whether the tests measure what they are intended to measure across different subgroups in the population. To the extent that the measurement model of the test that relates latent variables (e.g. reading skills) to observable indicators (e.g. test items) does not hold equivalently across different subgroups, measurement non-invariance is committed, which compromises test fairness (Liu & Dorans, 2016). Empirical evidence from the current project seems to indicate that measurement invariance is not maintained in the test. Multi-group analysis as imposed on the factor structure of the test (see chapter VI) suggested that the factor structure of the L-VSTEP reading test was non-invariant at the configural level across high- and low-achieving students,

and at the metric level across English pedagogy students, English for translation students, and non-English major students. More rigorous analysis at the metric level indicated that two higher-level subskills of understanding pragmatic meaning and summarizing textual information did not load equivalently onto the latent construct of reading ability across students with different academic disciplines. As revealed later in the semi-structured interviews with lecturers and students, these non-invariant loadings could be attributed to the unbalanced skill profiles required of students with different academic disciplines in the relevant academic programs. While students in the English major programs encounter the majority of reading subskills assessed in the test in their courses, non-English students were exposed to a limited range of reading subskills, with a particular focus on lower-level reading subskills only. These students, therefore, might have been placed at a disadvantage when it came to their performance on the test. This finding is worrying given that the test was intended to assess English reading ability for all students regardless of their academic disciplines. That non-English major students came to the test ill-prepared as opposed to their English major counterparts pointed to a huge gap in the English reading curricula and instructional practices for students across various disciplines. Unless more collaborative efforts are made among relevant stakeholders of the test to fill this gap, threats to the test fairness remain, which might seriously weaken the interpretation and use of its scores. The following implications, therefore, can be considered for curriculum designers, test makers, teachers, and policy makers.

The above findings have some implications for curriculum designers, researchers, teachers, students, and policy makers. As for curriculum designers and policy makers, collaborative efforts should be made to maintain the alignment between the English reading proficiency standards for graduation and the learning curricula. Meetings should be organized between these two key stakeholders so that a more informed and mutual understanding of the English requirements for tertiary students across different disciplines can be achieved. In addition, the English learning curricula should be updated regularly in response to the changing need of English reading proficiency so that what is taught and assessed reflects prospective social-economic situations. More collaborative efforts should also be made between the macro (e.g. policy makers and curriculum designers) and meso (e.g. teachers and students) level stakeholders so that requirements in terms of English proficiency, learning objectives, teaching and learning materials, methods and practices can be clearly articulated and voices of different stakeholders, particularly those at the meso level can be heard. These concerted efforts can be considered key to establishing and maintaining the alignment between the test and the learning curricula, thereby having the potential to provide a more transparent and meaningful

interpretation and use of the test scores. As for researchers, consistent research efforts should be made to probe further into the factors that may confound test fairness, particularly with regard to the interpretation of the test scores. Differential item functioning is another research area of test fairness that needs empirical justification for the test under investigation. The concept can be understood as the extent to which test items function differently for different subgroups of students who have the same level of proficiency but differ in the level of the respective construct-irrelevant factors, such as gender, socio-economic status, or academic learning experience (Aryadoust, 2013; Banerjee & Papageorgiou, 2016). The detection of differential item functioning, either using latent variable modelling or item response theory approaches, is useful to identify items that compromise test fairness and that need revision in the test development process. Finally, since English major and non-English major students are so fundamentally different in terms of the learning curriculum, English reading proficiency requirements, medium of instruction in class, and job-related English reading skills, there arises the need for developing two versions of the L-VSTEP reading, one for English major and one for non-English major students. These different versions of the test might better reflect the target language use domains and accommodate the interpretation and use of the test scores in the local context.

11.5. Limitations

Although the study offers useful implications for different stakeholders of the test, it is necessary to acknowledge the limitations that constrain the interpretation and generalization of the study's findings.

First, since the L-VSTEP test is a newly developed test of English proficiency in a local context in Vietnam, which remains subject to an ongoing development and validation process, and due to test score confidentiality policies, real data taken by test-takers sitting actual L-VSTEP tests were inaccessible. Simulation data in lieu of actual data, therefore, were used in the current project, which might not be as reliable as real data collected from students taking actual tests.

Second, several statistical methods (SEM, multigroup CFA, and Rasch) employed in the current project are based on large sample theory. Due to administrative restrictions and test material confidentiality policies, however, the number of available participants and test forms for some analyses was not sufficiently large, which implied that findings be interpreted with cautions. For example, the multiple regression analyses in Chapter VII were based on test items as units of analysis rather than on the participants who took the test. The use of test items as

units of analysis, however, was prone to the violation of the assumption of independence of observation (e.g. items are nested under paragraphs), which could have been better handled by other more advanced and large-sample based statistical modelling approaches, such as multi-level modeling (Hox, 2002; Luke, 2004). Similarly, since there was a large disproportion of male and female participants in the study, the multi-group analysis of the factor structure of the test across gender groups could not be conducted, which limited the generalizability of the study.

Third, since the study was situated in a local context, replication research attempts across the country would be needed to enhance the generalizability of the research findings. Finally, the current project focused only on the explanation and extrapolation inferences in Chapelle et al. (2008) argument-based approach. More similar validation studies would be needed to either explore other inferences in the argument-based framework, or to employ other validation frameworks such as Bachman & Palmer's (2010) assessment use argument framework or Weir's (2005) socio-cognitive framework to provide a more comprehensive account of the interpretation and use of the L-VSTEP reading test. For example, the utilization inference may need more attention in the validation of the interpretation and use of the L-VSTEP reading test as it is important to explore what attitude different stakeholders hold toward this newly-developed test.

11.6. Future directions

This study has employed multiple types of data to explore the technical issues of the test as well as the meaningfulness of the test scores beyond the test domain via the process of articulating and evaluating the interpretation and use argument. However, there are still other potential research areas that invite further investigation.

First, this study only involved students and teachers in generating data to answer the research questions. Voices of other relevant stakeholders, such as parents, policy makers, and curriculum designers would need to be explored in future studies to gain a more thorough understanding of the different stakeholders' perceptions about the interpretation and use of the test scores.

Second, the reading processes engaged by students in the current study were explored through the examination of stimulated verbal recall data. The only stimulus used to elicit readers' reading processes was their answers to the test items. Future studies that attempt to gain insights into students' reading processes are suggested to combine stimulated verbal recall with the use of eye tracking technology that has the potential to capture the second-by-second

movements of readers' eyes to provide on-the-spot, straightforward, and trustworthy evidence of the readers' actual reading behaviors.

Finally, this thesis has provided evidence in support of the use of self-assessment based on the CEFR-VN and the test development guidelines as a criterion measure for the L-VSTEP reading test. However, the accuracy of self-assessment, the various individual and experiential factors that may affect self-assessment accuracy, and the contexts in which self-assessment should be used to maximize its potential would need further empirical investigation so as to enable a more meaningful and effective use of self-assessment in language testing and assessment.

REFERENCES

- Akamatsu, N. (2003). The effects of first language orthographic features on second language reading in text. *Language learning*, 53(2), 207-231.
- Alderson, C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Alderson, C., & Lukmani, Y. (1989a). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5(2), 253-270.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: A&C Black.
- Alderson, J. C., & Lukmani, Y. (1989b). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5(2), 253-270.
- Alderson, J. C., & Lumley, T. (1995). Responses and replies. *Language Testing*, 12(1), 121-130.
- AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self-and peer-assessment. *System*, 32(3), 407-425.
- Alptekin, C. (2006). Cultural familiarity in inferential and literal comprehension in L2 reading. *System*, 34(4), 494-508.
- Alptekin, C., & Erçetin, G. (2011). Effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *Tesol Quarterly*, 45(2), 235-266.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, J. C. (1990). Testing reading comprehension skills (Part One). 6(2), 425-438.
- Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.
- Bamber, D., & van Santen, J. P. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, 29(4), 443-473.
- Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural equation modeling*, 9(1), 78-102.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307-337). Greenwich: Information Age Publishing
- Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, 30(1-2), 8-24.

- Barkaoui, K. (2015). *the characteristics of the Michigan English Test reading texts and items and their relationship to item difficulty*. Retrieved from <http://michiganassessment.org/wp-content/uploads/2015/04/CWP-2015-02.pdf>
- Barrot, J. S. (2016). ESL Learners' Use of Reading Strategies Across Different Text Types. *The Asia-Pacific Education Researcher*, 25(5-6), 883-892.
- Barton, M. E. (1994). *Input and interaction in language acquisition*. Cambridge: Cambridge University Press.
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441-465.
- Bax, S. (2015). *Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test*. Retrieved from https://www.ielts.org/-/media/research-reports/ielts_online_rr_2015-2.ashx
- Bax, S., & Weir, C. (2012). *Investigating learners' cognitive processes during a computer-based CAE Reading test*. Retrieved from <http://www.cambridgeenglish.org/images/22669-rv-research-notes-47.pdf>
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.
- Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing*, 37, 1-12.
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7(3), 303-310.
- Bell, L. C., & Perfetti, C. A. (1994). Reading skill: Some adult comparisons. *Journal of Educational Psychology*, 86(2), 244-255.
- Bensoussan, M. (1986). Beyond vocabulary: Pragmatic factors in reading comprehension—Culture, convention, coherence and cohesion. *Foreign Language Annals*, 19(5), 399-407.
- Bentler, P. M. (2008). *EQS program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16(1), 78-117.
- Best, R. M., Floyd, R. G., & Mcnamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading psychology*, 29(2), 137-164.
- Bloom, B. (1954). The thought processes of students in discussions. In S. J. French (Ed.), *Accent on teaching: experiments in general education* (pp. 23-46). New York, NY: Harper.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Understanding person measures. In W. J. Boone, J. R. Staver, & M. S. Yale (Eds.), *Rasch Analysis in the Human Sciences* (pp. 69-92). London, UK: Springer.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.
- Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System*, 36(3), 456-477.
- Brennan, R. L. (2001). *Generalizability*. New York, NY: Springer-Verlag.
- Brennan, R. L. (2013). Commentary on “validating the interpretations and uses of test scores”. *Journal of Educational Measurement*, 50(1), 74-83.
- British Council. (2017). Five facts on current education market in Vietnam. Retrieved from <https://ei.britishcouncil.org/news/5-facts-current-education-market-vietnam>

- Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly*, 11(4), 353-373.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals*, 47(2), 261-285.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.
- Brunfaut, T., & McCray, G. (2015a). Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study.
- Brunfaut, T., & McCray, G. (2015b). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*. Retrieved from https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf
- Buck, G., Tatsuoka, K., & Kostin, I. (1997). The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension. *Language learning*, 47(3), 423-466.
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.
- Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications and programming* (3rd ed.). London & New York: Routledge.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin*, 105(3), 456-466.
- Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, 23(3), 269-289.
- Carr, N. T., Nguyen, T. N. Q., Nguyen, T. M. H., Nguyen, T. Q. Y., Thai, H. L. T., & Nguyen, T. P. T. (2016). *Systematic support for a communicative standardized proficiency test in Vietnam*. Paper presented at the New directions in English language assessment, Hanoi. https://www.britishcouncil.vn/sites/default/files/new_directions_2016_nathan_carr_systematic_support_for_a_communicative_standardized_proficiency_test_in_vietnam.pdf
- Carr, T. H., & Levy, B. A. E. (1990). *Reading and its development: Component skills approaches*. San Diego, CA, US: Academic Press.
- Carrell, P. L. (1991). Second language reading: Reading ability or language proficiency? *Applied linguistics*, 12(2), 159-179.
- Carroll, B. J. (1980). *Testing communicative competence: An interim study*. Oxford: Pergamon Press.
- Carroll, J. B. (1993). *Human cognitive abilities: A study of factor-analytic studies*. Cambridge, UK: Cambridge University Press.
- Carver, R. P. (1992). Reading rate: theory, research, and practical implications. *Journal of Reading*, 36(2), 84-95.
- Chapelle, C. A. (2012a). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). New York, NY: Routledge.
- Chapelle, C. A. (2012b). Validity argument for language assessment: the framework is simple.... *Language Testing*, 29(1), 19-27.

- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3-13.
- Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-17). Chichester, UK: Wiley Blackwell.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling: A multidisciplinary journal*, 9(2), 233-255.
- Choi, I.-C., & Moon, Y. (2020). Predicting the Difficulty of EFL Tests Based on Corpus Linguistic Features and Expert Judgment. *Language Assessment Quarterly*, 17(1), 18-42.
- Chung, Y.-R. (2014). *A test of productive English grammatical ability in academic writing: development and validation*. (Doctoral dissertation). Iowa State University, Ames, IA.
- Cohen, A. D. (1984). On taking language tests: what the students report? *Language Testing*, 1(1), 70-81.
- Cohen, A. D. (2007). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307-331.
- Cohen, A. D. (2013). Using Test-Wiseness Strategy Research in Task Development. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 2, pp. 893-905). New Jersey: John Wiley & Sons Inc. .
- Cohen, A. D., & Macaro, E. (2007). *Language learner strategies*. Oxford, UK: Oxford University Press.
- Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. New Jersey: Wiley.
- Cohen, A. D., & Upton, T. A. (2007). I want to go back to the text!: response strategies on the reading subtest of the new TOEFL®. *Language Testing*, 24(2), 209-250.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Newbury Park, CA: Sage.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155-159.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4), 497-505.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Council of Europe. (2001). *Common European Framework of references for languages: learning, teaching, assessment* Cambridge, UK: Cambridge University Press.
- Creswell, J. W. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.
- Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage publications.
- Cronbach, L. J., & Meehl, P. E. (1955a). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281-302.
- Cronbach, L. J., & Meehl, P. E. (1955b). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281.
- Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: a case for an intuitive approach. *Language Teaching Research*, 16(1), 89-108.
- Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3), 475-493.

- Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15-30.
- Cunningham, A. E., Stanovich, K. E., & Wilson, M. R. (1990). Cognitive variation in adult college students differing in reading ability. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 129-159). San Diego, CA: Academic Press.
- Davis, F. (1968). Research in comprehension in reading. *Reading research quarterly*, 3(4), 499-545.
- Dawadi, S., & Shrestha, P. N. (2018). Construct validity of the Nepalese school leaving english reading test. *Educational Assessment*, 23(2), 102-120.
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (Fourth edition ed.). Thousand Oaks: Sage publications.
- Dolotic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living language: Self-assessment, oral production, and domestic immersion. *Foreign Language Annals*, 49(2), 302-316.
- Dornyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies* Oxford: Oxford University Press.
- Drackert, A. (2015). *Validating language proficiency assessments in second language acquisition*. Frankfurt: Peter Lang.
- Dunlea, J., Spiby, R., Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P. T., . . . Sao, B. T. (2018). *APTIS - VSTEP comparability study: investigating the usage of two EFL tests in the context of higher education in Vietnam*. Retrieved from <https://www.britishcouncil.org/aptis-vstep-comparability-study>
- Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychological bulletin*, 93(1), 179-197.
- Embretson, S. E., & Wetzel, D. C. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11(2), 175-193.
- Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework*. Princeton, NJ: Educational Testing Service.
- Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford: Oxford University Press.
- Fan, J. (2016). The construct and predictive validity of a self-assessment scale. *Papers in Language Testing and Assessment*, 5(2), 69-100.
- Fan, J., & Bond, T. G. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (pp. 83-102). New York, NY: Routledge.
- Fan, J., & Yan, X. (2017). From Test Performance to Language Use: Using Self-Assessment to Validate a High-Stakes English Proficiency Test. *The Asia-Pacific Education Researcher*, 26(1-2), 61-73.
- Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, 10(3), 274-291.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3), 209-226.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Field, A. (2009). *Discovering statistics using IBM SPSS statistics*. London, UK: Sage.

- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10(2), 133-170.
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1), 2-32.
- Frønes, T. S., Narvhus, E. K., & Aasebø, M. C. (2013). Nordic results from the PISA digital reading assessment. *Nordic Journal of Digital Literacy*, 8(01-02), 13-31.
- Fulcher, G. (1997). Text difficulty and reading formulae: Reading formulae and expert judgments. *System*, 25(4), 497-513.
- Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied linguistics*, 20(2), 221-236.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, UK: Routledge
- Fulcher, G., & Davidson, F. (2012). Introduction. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 1-19). London, UK: Routledge.
- Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28(1), 77-104.
- Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied linguistics*, 25(1), 1-37.
- Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*. London: Routledge.
- Geva, E., & Wang, M. (2001). The development of basic reading skills in children: A cross-language perspective. *Annual Review of Applied Linguistics*, 21(1), 182-204.
- Goodman, K. S. (1967). Reading: a psycholinguistic guessing game. *Literacy Research and Instruction*, 6(4), 126-135.
- Goodman, K. S. (1986). *What's whole in whole language? A parent/teacher guide to children's learning*. Portsmouth, NH: Heinemann.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5), 394-411.
- Goto Butler, Y., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27(1), 5-31.
- Gough, P. B. (1972). One second of reading. *Visible Language*, 6(4), 291-320.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education*, 7(1), 6-10.
- Government of Vietnam. (2008). Decree No. 1400/QĐ-TTg on the approval of the project on teaching and learning English in the national education system in the 2008 - 2020 period. Retrieved from http://www.chinhphu.vn/portal/page/portal/chinhphu/hethongvanban?class_id=1&_page=18&mode=detail&document_id=78437
- Grabe, W. P. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.
- Grabe, W. P., & Jiang, X. (2013). Assessing reading. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-16). Chichester, UK: Wiley Blackwell.
- Grabe, W. P., & Stoller, F. L. (2013). *Teaching and researching reading*. New York, NY: Routledge.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: providing multilevel analyses of text characteristics. *Educational researcher*, 40(5), 223-234.
- Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet

- & C. E. Snow (Eds.), *Rethinking reading comprehension* (Vol. 82-98, pp. 98). New York: Guilford Press.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2), 193-202.
- Grant, A., Gottardo, A., & Geva, E. (2011). Reading in English as a first or second language: The case of grade 3 Spanish, Portuguese, and English speakers. *Learning Disabilities Research & Practice*, 26(2), 67-83.
- Green, A. (1998). *Verbal protocol analysis in language testing research: a handbook*. Cambridge: Cambridge University Press.
- Green, A. (2014). *Exploring language assessment and testing: language in action*. London, UK: Routledge.
- Green, A., Ünalı, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191-211.
- Green, R. (2013). *Statistical analysis for language testers*. Hampshire: Palgrave Macmillan.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, 26(3), 499-510.
- Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, 31(1), 111-133.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement*. Westport, CT: American Council on Education/Praeger.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). London: Pearson.
- Halliday, M. (1978). *Language as social semiotic*. London, UK: Edward Arnold Publishers.
- Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.
- Hannon, B. (2012). Understanding the relative contributions of lower-level word processes, higher-level processes, and working memory to reading comprehension performance in proficient adult readers. *Reading research quarterly*, 47(2), 125-152.
- Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, 93(1), 103-128.
- Heaton, J. B. (1975). *Writing English language tests: a practical guide for teachers of English as a second or foreign language*. London, UK: Longman.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, Massachusetts: Newbury House Publishers.
- Ho, R. (2014). *Handbook of univariate and multivariate data analysis with IBM SPSS*. Danver, MA: CRC Press.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2(2), 127-160.
- Horiba, Y., & Fukaya, K. (2015). Reading and learning from L2 text: Effects of reading goal, topic familiarity, and language proficiency. *Reading in a Foreign Language*, 27(1), 22-46.
- Hox, J. J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1-55.
- Huang, H.-T. D. (2016). Exploring strategy use in L2 speaking assessment. *System*, 63, 13-27.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

- In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, 26(2), 219-244.
- In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, 8(3), 250-276.
- In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, 29(1), 131-152.
- International Language Testing Association. (2007). ILTA guidelines for practice. Retrieved from https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/ilta_guidelines.pdf
- Jackson, N. E. (2005). Are university students' component reading skills related to their text comprehension and academic achievement? *Learning and Individual Differences*, 15(2), 113-139.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, 26(1), 31-73.
- Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics*, 34, 123-153.
- Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. New York: Routledge.
- Jia, Y. (2013). *Justifying the use of a second language oral test as an exit test in Hong Kong: An application of assessment use argument framework*. Doctoral dissertation. University of California, LA.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14-26.
- Johnson, R. C., & Riazi, A. M. (2015). Accuplacer Companion in a foreign language context: An argument-based validation of both test score meaning and impact. *Papers in Language Testing and Assessment*, 4(1), 31-58.
- Johnson, R. C., & Riazi, A. M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing*, 32, 85-104.
- Jun, H. S. (2014). *A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test*. Doctoral dissertation. Iowa State University. Ames, IA.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1), 122-149.
- Kadir, K. A. (2008). *Framing a validity argument for test use and impact: The Malaysian public service experience*. Doctoral dissertation. University of Illinois at Urbana-Champaign, IL.
- Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, 35(4), 401-415.
- Kane, M. (1992). An argument-based approach to validity. *Psychological bulletin*, 112(3), 527-535.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2(3), 135-170.
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3-17.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, 18(2), 5-17.

- Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
- Kim, A.-Y. (2009). Investigating Second Language Reading Components: Reading for Different Types of Meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 9(2), 1-28.
- Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32(2), 227-258.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.
- Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275-305). New York, NY: Routledge.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193-220.
- Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, 1(1), 60-69.
- Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language testing in Asia*, 6(3), 1-23.
- Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27(2), 183-189.
- LaBerge, D., & Samuels, J. S. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, 6(2), 293-323.
- Lado, R. (1961). *Language Testing: the Construction and Use of Foreign Language Tests*. New York: McGraw - Hill
- LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, 34(4), 451-475.
- Lallmamode, S. P., Daud, N. M., & Kassim, N. L. A. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing*, 30, 44-62.
- Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing*, 23(6), 701-717.
- Landi, N., & Perfetti, C. A. (2007). An electrophysiological investigation of semantic and phonological processing in skilled and less-skilled comprehenders. *Brain and language*, 102(1), 30-45.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Béjoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Springer.
- Laufer, B. (1996). The lexical threshold of second language reading comprehension: What it is and how it relates to L1 reading ability. In K. Sajavaara & C. Fairweather (Eds.), *Approaches to second language acquisition* (pp. 55-62). Yvaskyla, Finland: University of Jyvaskyla.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, 16(3), 307-322.
- Le, V. C. (2017). English language education in Vietnamese universities: National benchmarking in practice. In B. Spolsky, H. Terauchi, & W. K. Too (Eds.), *English education at the tertiary level in Asia: From policy to practice* (pp. 183-203). New York: Routledge.

- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.
- Li, Z. (2013). The issues of construct definition and assessment authenticity in videobased listening comprehension tests: Using an argument-based validation approach. *International Journal of Language Studies*, 7(2), 61-82.
- Li, Z. (2015a). *An argument-based validation study of the English Placement Test (EPT) – Focusing on the inferences of extrapolation and ramification*. (Doctoral dissertation). Iowa State University, Retrieved from <https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=5545&context=etd>
- Li, Z. (2015b). Using an English self-assessment tool to validate an English Placement Test. *Papers in Language Testing and Assessment*, 4(1), 59-96.
- Linacre, J. M. (2014). *Winstep Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winstep.com.
- Linacre, J. M. (2020). *A user's guide to WINSTEPS/MINISTEP: Rasch model computer program*. Chicago, IL: Winsteps.
- Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, 94(4), 778-784.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural equation modeling*, 9(2), 151-173.
- Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System*, 80, 60-72.
- Liu, J., & Dorans, N. J. (2016). Fairness in score interpretation In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 77-96). New York: Routledge.
- Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, 27(3), 32-42.
- Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182-212). London: Routledge.
- Louwerse, M. (2001). An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics*, 12(3), 291-316.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *Tesol Quarterly*, 45(1), 36-62.
- Luke, D. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.
- Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211-234.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130-149.
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85-104.
- Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, 2(4), 260-293.
- McCarthy, P. M., & Jarvis, S. (2010). MTL, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392.

- McCormick, S., & Zutell, J. (2007). *Instructing students who have literacy problems* (5th ed.). Upper Saddle River, NJ: Pearson Education.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, 27(1), 57-86.
- Mcnamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.
- McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, 8(2), 161-178.
- Menard, S. (1995). *Applied logistic regression analysis* (Vol. 106). Thousand Oaks: Sage.
- Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, 35, 41-55.
- Messick. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York, NY: Macmillan Publishing.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International journal of lexicography*, 3(4), 235-244.
- Ministry of Education and Training of Vietnam. (2015a). *Decision No. 730/QĐ-BGDĐT on the approval of the guidelines for the development and marking of the test of English proficiency at levels 3 to 5 of the CEFR-VN for Vietnamese learners of English*.
- Ministry of Education and Training of Vietnam. (2015b). Decree No. 729/QĐ-BGDĐT on the circulation of the test format for the Vietnamese Standardized Test of English Proficiency targeted at levels 3 - 5 of the 6-level National Standard of language proficiency. Retrieved from <https://thuvienphapluat.vn/van-ban/Giao-duc/Quy-et-dinh-729-QD-BGDĐT-2015-de-thi-danh-gia-nang-luc-su-dung-tieng-Anh-tu-bac-3-den-bac-5-267956.aspx>
- Ministry of Education and Training of Vietnam (MOET). (2014). Circular No. 01/2014/TT-BGDĐT on the 6-level Vietnamese National Standard of Language Proficiency. Retrieved from <https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-01-2014-TT-BGDĐT-Khung-nang-luc-ngoai-ngu-6-bac-Viet-Nam-220349.aspx>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1), 3-62.
- Moeller, A. J., Creswell, J. W., & Saville, N. (2016). *Second language assessment and mixed methods research*. Cambridge, UK: Cambridge University Press.
- Munby, J. L. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.
- Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Belmont, CA: Duxbury Press
- Myers, S. S. (1991). Performance in reading comprehension— product or process? *Educational Review*, 43(3), 257-272.
- Nassaji, H. (2002). Schema Theory and Knowledge-Based Processes in Second Language Reading Comprehension: A Need for Alternative Perspectives. *Language learning*, 52(2), 439-481.
- Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal*, 87(2), 261-276.
- Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal*, 90(3), 387-401.

- Nassaji, H. (2014). The role and importance of lower-level processes in second language reading. *Language Teaching*, 47(1), 1-37.
- Nation, I. (1993). Vocabulary size, growth, and use. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (Vol. 6, pp. 115-134). Amsterdam: John Benjamins.
- Nation, K., & Snowling, M. J. (1998). Semantic processing and the development of word-recognition skills: evidence from children with reading comprehension difficulties. *Journal of memory and language*, 39(1), 85-101.
- Nguyen, T. N. Q. (2020). Vietnamese Standardized Test of English Proficiency: A panorama. In L. I.-W. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English language proficiency testing in Asia* (pp. 71-100). New York: Routledge.
- Nguyen, T. P. T. (2018). An investigation into the content validity of a Vietnamese standardized test of English proficiency (VSTEP. 3-5) reading test. *VNU Journal of Foreign Studies*, 34(4), 129-143.
- Nguyen, V. H., & Hamid, M. O. (2015). Educational policy borrowing in a globalized world. *English Teaching: Practice & Critique*, 14(1), 60-74.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3), 231-259.
- O'Rourke, N., Hatcher, L., & Stepanski, E. J. (2005). *A step-by-step approach to using SAS for univariate and multivariate statistics* (2nd ed.). Cary, NC: SAS Institute Inc.
- Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, 12(3), 305-319.
- Oller, J. W. (1979). *Language tests at school*. London, UK: Longman.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, 24(4), 492-518.
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. *Encyclopedia of language and education*, 7, 175-187.
- Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: the passage or the question? *Behavior Research Methods*, 40(4), 1001-1015.
- Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, 7(2), 137-159.
- Perfetti, C. A. (1985). *Reading ability*. Oxford: Oxford University Press.
- Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G. E. Mackinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (pp. 109-143). San Diego, CA: Academic Press.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357-383.
- Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. Just & P. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 2-65). Hillsdale, MI: Lawrence Erlbaum Associates.
- Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury.
- Phakiti, A., & Roever, C. (2018). *Quantitative methods for second language research*. New York, NY: Routledge.
- Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC® Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing*, 32(2), 151-167.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41, 71-90.

- Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8-25.
- Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes, 20*, 10-16.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin, 114*(3), 552-566.
- Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics*. London: Routledge.
- Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: opportunities, issues and challenges. *Language Teaching, 47*(2), 135-173.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20.
- Rupp, A. A. (2012). Psychological versus psychometric dimensionality in reading assessment. In J. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 135-153). New York, NY: Rowan & Littlefield Education.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing, 23*(4), 441-474.
- Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1*(3-4), 185-216.
- Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & cognition, 17*(6), 759-769.
- Saadatnia, M., Ketabi, S., & Tavakoli, M. (2017). Levels of reading comprehension across text types: A comparison of literal and inferential comprehension of expository and narrative texts in Iranian EFL learners. *Journal of psycholinguistic research, 46*(5), 1087-1099.
- Sadighi, S., Yamini, M., Bagheri, M. S., & Yarmohammadi, L. (2018). Investigating preuniversity EFL teachers' perceived wash-back effects of university entrance exams and teaching materials on students' learning objectives and teachers' class performance. *Cogent Social Sciences, 4*(1), 1-16.
- Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993a). A causal model of sentence recall: effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of reading behavior, 25*(1), 5-16.
- Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993b). Impact of concreteness on comprehensibility, interest, and memory for text: implications for dual coding theory and text design. *Journal of Educational Psychology, 85*(2), 291-304.
- Sadoski, M., Goetz, E. T., & Rodriguez, M. (2000). Engaging texts: effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology, 92*(1), 85-95.
- Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing*. London, UK: Routledge.
- Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In v. A. Eye & C. C. Clogg (Eds.), *Latent variables analysis: applications for developmental research*. Thousand Oaks, CA: Sage.
- Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing, 35*(4), 529-556.

- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 5-30.
- Schedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1995). An analysis of the dimensionality of TOEFL reading comprehension items. *ETS Research Report Series*, 1995(2), 1-26.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: detection, search, and attention. *Psychological review*, 84(1), 1-66.
- Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York, NY: Routledge.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4), 523-568.
- Shankweiler, D., Lundquist, E., Dreyer, L. G., & Dickinson, C. C. (1996). Reading and spelling difficulties in high school students: causes and consequences. *Reading and Writing*, 8(3), 267-294.
- Sheehan, K., & Ginther, A. (2001). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*. Paper presented at the Annual meeting of the National Council of Measurement in Education, Seattle, WA.
- Shiotsu, T. (2010). *Components of L2 reading: linguistic and processing factors in the reading test performances of Japanese EFL learners* (Vol. 32). Cambridge, UK: Cambridge University Press.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1(2), 147-170.
- Snowling, M., Hulme, C., & Nation, K. (1997). A connectionist perspective on the development of reading skills in children. *Trends in cognitive sciences*, 1(3), 88-91.
- Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435-464.
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly*, 32-71.
- Stanovich, K. E. (1982). Individual differences in the cognitive processes of reading: word decoding. *Journal of Learning Disabilities*, 15(8), 485-493.
- Stanovich, K. E. (1984). The interactive-compensatory model of reading: a confluence of developmental, experimental, and educational psychology. *Remedial and special education*, 5(3), 11-19.
- Stanovich, K. E., West, R. F., & Cunningham, A. E. (1991). Beyond phonological processes: print exposure and orthographic processing. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: a tribute to Isabelle Y. Liberman* (pp. 219-235). Mahwah, NJ: Routledge.
- Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, 32(1), 63-81.
- Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping TOEIC scores to the Vietnamese National Standard: a study to recommend English language requirements for admission into and graduation from Vietnamese universities*. Princeton, NJ: Educational Testing Service.
- Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 3-50). Thousand Oaks: Sage.
- Tengberg, M. (2018). Validation of sub-constructs in reading comprehension tests using teachers' classification of cognitive targets. *Language Assessment Quarterly*, 15(2), 169-183.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

- Toulmin, S. E. (2003). *The uses of arguments* (2nd ed.). Cambridge, UK: Cambridge University Press.
- University of Languages and International Studies. (2015). *Format and samples of the VSTEP test*. Hanoi: National University of Hanoi Publishing House.
- Upshur, J. (1975). *Objective evaluation of oral proficiency*. Paper presented at the English Training Forum.
- Urquhart, S., & Weir, C. (1998). *Reading in a second language: Process, product and practice*. London, UK: Routledge.
- Van Steensel, R., Oostdam, R., & Van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing*, 30(1), 3-21.
- Verhoeven, L., & van Leeuwe, J. (2009). Modeling the growth of word-decoding skills: Evidence from Dutch. *Scientific Studies of Reading*, 13(3), 205-223.
- Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design*. Doctoral dissertation. Iowa State University. Ames, IA.
- Walczyk, J. J. (1995). Testing a compensatory-encoding model. *Reading research quarterly*, 30(3), 396-408.
- Walczyk, J. J. (2000). The interplay between automatic and control processes in reading. *Reading research quarterly*, 35(4), 554-566.
- Wang, D. (2009). Factors affecting the comprehension of global and local main idea. *Journal of College Reading and Learning*, 39(2), 34-52.
- Wang, J., & Wang, X. (2012). *Structural equation modeling: applications using Mplus*. Chichester, UK: John Wiley & Son Ltd.
- Weir, C. (2005). *Language testing and validation*. Hampshire: Palgrave MacMillan.
- Weir, C., Hawkey, R., Green, A., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume09_report4.ashx
- Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL: Research Notes*, 31, 2-10.
- Weir, C., & Porter, D. (1994). The multi-divisible or unitary nature of reading: the language tester between scylla and charybdis. *Reading in a Foreign Language*, 10(2), 1-19.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56-75). Newbery Park, CA: SAGE.
- Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford, UK: Oxford University Press.
- Wright, B., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions*, 9(4), 468-468.
- Yamashita, J. (2002). Influence of L1 reading on L2 reading: Different perspectives from the process and product of reading. *Studies in Language and Culture*, 23(2), 271-283.
- Yan, X., & Staples, S. (2020). Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE. *Language Testing*, 37(2), 189-214.
- Yang, H. J. (2016). *Integration of a web-based rating system with an oral proficiency interview test: argument-based approach to validation*. (Doctoral). Iowa State University, Iowa.
- Yoo, H., & Manna, V. F. (2017). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing*, 34(1), 101-126.

- Yoshida, M. (2012). The interplay of processing task, text type, and proficiency in L2 reading. *Reading in a Foreign Language, 24*(1), 1-29.
- Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika, 67*(1), 95-121.
- Zheng, Y., Cheng, L., & Klinger, D. A. (2007). Do test formats in reading comprehension affect second-language students' test performance differently? *TESL Canada Journal, 25*(1), 65-80.

APPENDICES

Appendix 1: Ethics approval


Ethics application - approved - 1800001029

Research Ethics (HUMAN) <humanethics@qut.edu.au>

Mon 12/10/2018 8:39 AM

To: Lynette May <lynette.may@qut.edu.au>; Michael Mu <m.mu@qut.edu.au>; Ngoc Hoi Vo <hoi.vongoc@hdr.qut.edu.au>

Cc: Human Ethics Advisory Team <humanethics@qut.edu.au>

 1 attachments (119 KB)

RESEARCHGOVCHKLIST_20180704.PDF;

Dear Dr Lynette May and Mr Vo Ngoc Hoi

Ethics Category: Human - Negligible-Low Risk

UHREC Reference number: 1800001029

Dates of approval: 10/12/2018 to 10/12/2023

Project title: Validity evidence of the

Vietnamese standardised test of English proficiency reading component: An argument-based approach

Thank you for submitting the above research project for ethics review.

This project was considered by Chair, Queensland University of Technology (QUT) Human Research Ethics Committee (UHREC) or a Faculty-based low risk review panel.

We are pleased to advise you that the above research project meets the requirements of the National Statement on Ethical Conduct in Human Research (2007) and ethics approval for this research project has been granted on behalf of the UHREC, to be ratified at their next scheduled meeting.

Appendix 2: The expert judgment form

<i>Items</i>	Understanding explicit information		Understanding cohesive devices		Integrating textual information		Summarizing textual information		Inferring situational meaning		Understanding pragmatic meaning		Lexical inferencing		Identifying text structure		Other skills	
	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>	<i>Pri</i>	<i>Sec</i>
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		
25																		
26																		
27																		

28																		
29																		
30																		
31																		
32																		
33																		
34																		
35																		
36																		
37																		
38																		
39																		
40																		

Appendix 3: Reading skill definition, description, and sample questions

Skill	Definition	Description	Sample question
Understanding explicit information at local level	- The ability to locate and understand explicit meaning at the sentence level.	<ul style="list-style-type: none"> - Understand specific details that are explicitly stated in the texts, using simple grammatical structures and vocabulary - Locate a specific detail in the text. - Identify and understand paraphrased information explicitly stated in the text. 	<p>On average, how much do tenants have to pay for a studio in New York City?</p> <p>A. About \$2,000</p> <p>B. More than \$2,000</p> <p>C. More than \$3,100</p> <p>D. Less than \$3,500</p>
Understanding cohesive devices	- The ability to understand the relationship between sentences or ideas using connective devices such as discourse markers, anaphoric and cataphoric references, substitutions, repetitions.	<ul style="list-style-type: none"> - Identify the antecedent of a pronoun. - Understand logical ideas in the text based on linking devices such as referent words, conjunctions, linking words, and repeated words. 	<p>The word ‘they’ in line 3 refers to</p> <p>A. mats</p> <p>B. origins</p> <p>C. bacteria</p> <p>D. DNA</p>
Integrating textual information	- The ability to synthesize information from different parts of a paragraph or a text.	<ul style="list-style-type: none"> - Locate and synthesize information across a paragraph. - Locate and synthesize information across the text. 	<p>According to the passage, who is likely to meet different types of people every day?</p> <p>A. Luc</p> <p>B. Harry</p> <p>C. Jennifer</p> <p>D. Solange</p>
Summarizing textual information	- The ability to understand main ideas and recognize supporting	- Understand the main idea of a paragraph.	What is the main idea of paragraph 3?

	details at paragraph and discourse level.	<ul style="list-style-type: none"> - Understand the main idea of a text - Identify and understand supporting details for an argument or the main ideas of a paragraph or a text. 	<p>A. Hot weather combined with wild fire soot has been melting glaciers.</p> <p>B. There has been enough evidence that global warming is an urgent issue.</p> <p>C. Global warming is evident but some are not willing to deal with this.</p> <p>D. The earliest effects of melting glaciers can only be seen in centuries.</p>
Inferring situational meaning	- The ability to make inferences about details, relationships, situations, and arguments using textual or background knowledge.	<ul style="list-style-type: none"> - Identify and understand an implicit detail that is rewritten using different words. - Understand the underlying meaning of a sentence or a detail. - Understand the logical inference of an argument. 	<p>What does Robert Shapiro mean when he says, ‘To adopt this, you have to believe we were incredibly lucky’?</p> <p>A. Supporters of RNA world hypothesis must think that humans were extremely blessed.</p> <p>B. Humans were incredibly lucky because the RNA was the first form of life on Earth.</p> <p>C. He believes it is near impossible that RNA accidentally arose on Earth.</p> <p>D. Humans were unlucky because the RNA world hypothesis is highly improbable.</p>
Understanding pragmatic meaning	- The ability to understand author’s purpose, attitude, tone, mood, belief, and intention in the text.	<ul style="list-style-type: none"> - Understand the author’s purpose, attitude, opinion, or stance on an issue in the text. - Understand the general tone of a text. - Understand the purpose of the author via a detail in the text. 	<p>What is the author’s purpose when recounting the scene he saw from the plane?</p> <p>A. To introduce the idea of global warming</p> <p>B. To give specific detail to support his point that global warming needs public awareness</p>

			<p>C. To express his opinion towards research on global surface temperature</p> <p>D. To contrast with what the pilot is saying</p>
Lexical inferencing	- The ability to guess the meaning of words using contextual clues.	<p>- Guessing word meaning from contexts (words with different meanings)</p> <p>- Guessing word meanings from contexts (idiomatic expressions)</p>	<p>What is ‘offer comfort’ in line 16 closest in meaning to?</p> <p>A. Warm up</p> <p>B. Reassure</p> <p>C. Discourage</p> <p>D. Assist</p>
Identifying genre and text structure at discourse level	- The ability to identify the genre (such as narrative, expository, persuasive, joke, diary) or identify structure of information and ideas at the discourse level (such as problem – solution, cause – effect, comparison, contrast).	<p>- Identify the organizational structure of a text.</p> <p>- Identify the genre of a text.</p>	<p>Which of the following best describes the organization of this passage?</p> <p>A. A general presentation followed by a detailed discussion of both sides of an issue.</p> <p>B. A list of possible answers to a question followed by a discussion of their strengths and weaknesses.</p> <p>C. A general statement of an issue followed by a discussion of possible answers. D. A discussion of different aspects wrapped up by an answer to the question.</p>

Appendix 4: The English reading proficiency self-assessment questionnaire

THE ENGLISH READING PROFICIENCY SELF-ASSESSMENT QUESTIONNAIRE

Instruction: In this section, you are asked to think about your English reading performance both in and outside the class throughout your tertiary learning so far and respond to the following statements. Please indicate the degree to which you agree with each statement by ticking in the relevant box.

Statements	Strongly disagree	Disagree	Somewhat disagree	Somewhat agree	Agree	Strongly agree
1. I can understand factual information in a text, using grammatical structures and vocabulary familiar to me.						
2. I can understand an explicit detail in the text but rewritten using different words.						
3. I can understand concepts/ideas in sentences that use familiar grammatical structures and vocabulary, have familiar topics and clear organization.						
4. I can identify the perspectives and stances of the text's author.						
5. I can understand the tone of the author in a text.						
6. I can identify the message that the author						

wants to convey via the text.						
7. I can understand the purpose of the author given a detail in the text.						
8. I can understand the logical inferences/arguments of the author.						
9. I can identify the antecedent of a pronoun.						
10. I can understand the logical ideas among sentences in a text, based on reference words, linking words, connectives, or repeated words.						
11. I can understand the main idea of a whole text.						
12. I can understand the main idea of a paragraph.						
13. I can draw the conclusions from a passage.						
14. I can integrate information across the text to establish the logical connection among ideas.						
15. I can integrate information in the text						

with my prior knowledge to understand an argument/detail.						
16. I can infer the meaning of colloquial or idiomatic expressions from the context.						
17. I can infer the implied meaning of a detail based on information within the text.						
18. I can infer the implied meaning of an argument in the text.						
19. I can infer the meaning of a subtle detail about an opinion/inference/attitude in a text.						
20. I can infer the implied meaning of a sentence/detail in a text using my prior knowledge.						
21. I can infer word meanings from contexts (words with different meanings)						
22. I can understand the supporting details for the main idea of a paragraph						

23. I can locate the specific information in a text.						
24. I can understand the function of a text.						
25. I can identify the organizational structure of a text (e.g. cause-effect, problem-solution)						
26. I can understand key ideas in a text.						
27. I can understand implicit information rewritten using different words.						

THANK YOU VERY MUCH FOR YOUR COOPERATION!